The Placebo Effect of Artificial Intelligence in Human-Computer Interaction

THOMAS KOSCH*, Utrecht University, the Netherlands

ROBIN WELSCH*, LMU Munich, Germany

LEWIS CHUANG, Chemnitz University of Technology, Germany

ALBRECHT SCHMIDT, LMU Munich, Germany

In medicine, patients can obtain real benefits from a sham treatment. These benefits are known as the placebo effect. We report two experiments (Experiment I: N=369; Experiment II: N=100) demonstrating a placebo effect in adaptive interfaces. Participants were asked to solve word puzzles while being supported by no system or an adaptive AI interface. All participants experienced the same word puzzle difficulty and had no support from an AI throughout the experiments. Our results showed that the belief of receiving adaptive AI support increases expectations regarding the participant's own task performance, sustained after interaction. These expectations were positively correlated to performance, as indicated by the number of solved word puzzles. We integrate our findings into technological acceptance theories and discuss implications for the future assessment of AI-based user interfaces and novel technologies. We argue that system descriptions can elicit placebo effects through user expectations biasing the results of user-centered studies.

CCS Concepts: • **Human-centered computing** → **User studies**; *HCI theory, concepts and models*; *Empirical studies in HCI*.

Additional Key Words and Phrases: Placebo Effect; Human-AI Interfaces; User Expectations; Placebo; User Studies

1 INTRODUCTION

A medical placebo; a sham treatment (e.g., a sugar pill without active ingredients), can improve a patient's subjective condition [29, 44] without treating the patient with an active substance or specific procedure. Such a placebo can alleviate pain [60] or support the treatment of ailments [3, 48], thereby providing an effective medical treatment without an illness-specific mechanism of effect. The only critical determinant of the placebo effect is the patient's expectation in the placebo's efficacy, leading to a positive evaluation after treatment [72, 95].

Regardless of its usefulness to treat ailments, the placebo effect obscures the evaluation of novel medical interventions. A novel cure could be evaluated positively only due to the expectation of improvement. Therefore, medical studies employ placebo-control when evaluating a novel treatment. Treatments are deemed effective only if their effect is larger than a placebo-control. The improvement caused by a specific treatment has to exceed the improvement brought about by expectations alone. This placebo-control is widely implemented in other scientific fields that evaluate human responses, such as psychological treatment [6], sport science [77], and even visualization research [12]. However, it is not the norm when evaluating the effectiveness of novel human-machine interfaces.

---

*Contributed equally to this research.

Authors' addresses: Thomas Kosch, Utrecht University, Utrecht, the Netherlands, t.a.kosch@uu.nl; Robin Welsch, LMU Munich, Munich, Germany, robin.welsch@ifi.lmu.de; Lewis Chuang, Chemnitz University of Technology , Chemnitz, Germany, clew@hrz.tu-chemnitz.de; Albrecht Schmidt, LMU Munich, Munich, Germany, albrecht.schmidt@ifi.lmu.de.
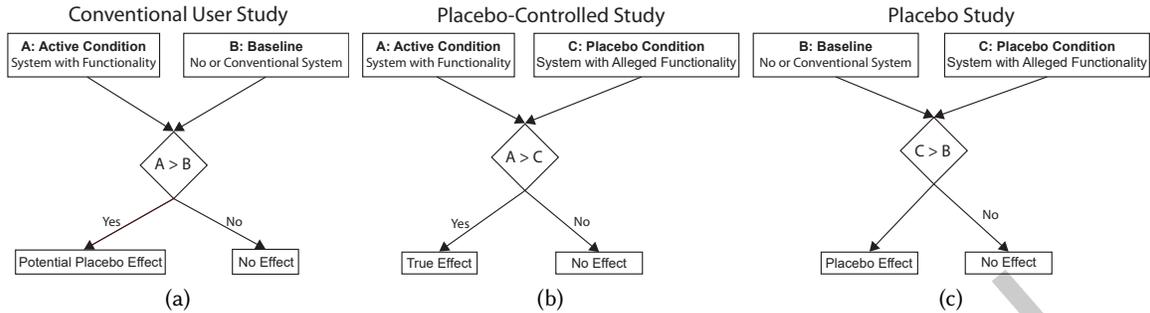
Fig. 1. Flow charts comparing conventional and placebo-controlled studies. **(a):** Conventional user studies comparing novel systems with a baseline do not take the potential presence of a placebo into account, which may be responsible for the perceived superiority of the novel system. **(b):** Instead, comparing a novel system with a system that pretends an alleged novel functionality does reveal if the actual novel system has an effect. **(c):** A placebo effect in the study measures is present if the alleged novel system is considered superior to a conventional baseline.

What would the placebo effect in Human-Computer Interaction (HCI) look like? Novel technological systems are evaluated by presenting systems to users and measuring their responses (see Figure 1a). A placebo effect in HCI would consist of a system to be judged as more usable or to return better task performance after interaction because participants have heightened expectations in the system's capabilities prior to the interaction. In medicine, expectations of improvement are controlled for by giving a non-active substance to participants who are led to believe that the substance is active and will have an effect. The participants will not know whether they have been given the active treatment (e.g. a painkiller) or an inactive substance (e.g. a sugar pill). In HCI, participants often recognize the system or technique and can tell apart the experimental conditions, e.g. a novel AI-system against a baseline system (see Figure 1a). Instructions might even explicitly describe the novelty of the evaluated systems and, hence, set participants' expectations. Thus, a placebo effect in HCI would consist of a participant's favorable evaluation in terms of effectiveness after the interaction. This includes an improvement of task performance in the absence of an active system (e.g., thinking an AI would support users even when *No Adaptivity* is given, see Figure 1c). The current work investigates to what extent descriptions of systems can alter subjective evaluations and task performance in user studies.

Previous HCI studies have demonstrated that a placebo can improve user experience. In games, fake power-up elements that make no difference to gameplay [20], and sham descriptions of Artificial Intelligence (AI) adaptation increase the self-reported game immersion [18]. In social media, providing control settings for prioritizing items in one's news feed can result in higher subjective ratings of user satisfaction, even when these control settings do not influence anything [100]. Sham design elements thus affect the pleasantness of the experience with interactive technologies. We are unaware, however, of any evidence for a placebo effect in HCI on personal effectiveness after interaction or objective task performance, similar to placebos in medical trials.

The current work investigates if a sham treatment (i.e., a placebo), as operationalized by system descriptions, can influence subjective and objective metrics of effectiveness when engaging with a system. Drawing from placebo-theory in psychology and medicine [3, 6, 60, 95], we systematically list HCI studies controlling for placebo effects. We then provide an literature search on the consideration of placebo effects in HCI. Motivated by the classification of the search results (see Figure 1), we ran two experiments manipulating system descriptions [24] aligning with placebo-theory [6, 29, 95] (see Figure 1c).

We provided participants with different descriptions of an adaptive AI that supports the completion of a word puzzle task following the example of previous studies [18, 20, 100]. We compare the participants' evaluations and task performance to control conditions. We realized this by instructing participants that a novel AI-based system would analyze their webcam image and choose word puzzles (see Figure 2), varying in difficulty to optimize their task performance. In reality, all participants underwent the same procedure — *No Adaptivity* was always provided where no adaptation took place. Note that different definitions of AI exist (for a range of potential definitions and conceptualizations, see [40, 63, 79]). Here, we see AI as a machine learning-based algorithm that is capable of processing and integrating complex data (e.g. a video feed) to make decisions. Input and output resemble human decision-making (e.g. seeing arousal in the face of a participant and giving an easier item). In our experiments, only the implementation of expectations by our system descriptions and not the concept of adaptation was critical.

In two online experiments, we investigated the expectancy-placebo effect of system descriptions for a between-subjects (i.e. Experiment I) and within-subjects (i.e. Experiment II) design. In Experiment I, we tested two versions of an adaptive AI system against a *No Adaptivity*-condition. Thus, a third of participants received a description of a *Physiology-based Adaptation*, AI-based assistance system adjusting the word puzzle difficulty using the video feed of the participant webcam. Another third received a description of an *Error-based Adaptation*, where the word puzzle difficulty was adapted based on task errors. Here, erroneous responses would simplify the following questions and correct responses increase the difficulty. The final third of participants were told that they would not receive any support, hence receiving *No Adaptivity* at all. While the *Error-based Adaptation* condition presents a control condition of an AI system description in which participants could validate the adaptive system, the *Physiology-based Adaptation* presents a placebo condition — a description of an AI-based system that has been employed in several HCI studies [18, 20, 100]. Experiment II compared a *Physiology-based Adaptation* to a *No Adaptivity* condition in a more controlled within-subjects study.

Experiment I found increased user expectations for *Physiology-based Adaptation* and *Error-based Adaptation* system descriptions. However, we did not find a significant effect between the system descriptions and the error rate, showing that the objective performance was not impacted using different system narratives. To investigate further if individual expectations are responsible for this, Experiment II used a within-subject design to control for potential subjective differences. Here, we utilized the *No Adaptivity* and *Physiology-based Adaptation* condition, showing a placebo effect on the subjective performance ratings when using *Physiology-based Adaptation*. Again, no statistical significant effect was found for the error rate, suggesting that placebo effects of AI systems do not affect objective measures but are instead impacting subjective measures.

CONTRIBUTION STATEMENT

This paper makes two contributions: Inspired by a literature search on placebo elements in HCI studies, we demonstrate a placebo effect in HCI in two experiments: (1) In a word puzzle task, we present participants with three different system descriptions suggesting improvements in the subjective performance ratings after interaction. (2) We then discuss the implications of a placebo effect in evaluation studies of AI and other systems and show how user expectations can be assessed and controlled for between- and within-subject experimental designs.

2 RELATED WORK

First, we present standard measures that assess the usability and acceptance of novel user interfaces. Next, we describe research on adaptive AI-based interfaces. Finally, we provide a systemic literature survey summarizing research in the domain of placebo effects in HCI.

2.1 User Expectations

User expectations concerning satisfaction are particularly relevant when evaluating novel technology. If interaction matches the user's prior expectations, user interfaces are typically considered more favorably [15, 98, 103] (e.g., an adaptive system supporting task completion) in terms of user satisfaction. Disconfirmed expectations (e.g., when an innovative technology makes task completion harder or does not add support), can produce adverse effects concerning user satisfaction and can be the primary factor in the rejection of a novel technology [65]. This has been shown in survey-based methods [15, 98, 103] and also experimentally [4, 39, 78]. In a gaming context, Michalco et al. [69] primed users with positive, neutral, or negative game ratings to investigate how knowledge about ratings influences the user's assessment of the game. Their results show that a positive prime can foster more positive ratings after playing the game if user expectations are confirmed, but also that dis-confirmation of expectation leads to even stronger contrasting effects on ratings, i.e., positively primed. Still, unenjoyable games are rated to be more negative. Therefore, providing users with prior knowledge on the satisfaction of others can significantly influence user evaluation. In line with this, Hollis et al. [41] and Costa et al. [13] showed that providing users with made-up positive feedback on their physiological state (e.g. electrodermal activity or heart rate) can influence their subjective emotional state [41] and can be used to reduce feelings of anxiety [13]. The focus of these studies was on user satisfaction with the interactive technology or well-being in response to fake feedback. In our study, we implemented expectations with regard to the effectiveness of the system that if sustained after interaction can be considered placebo effects. Although user satisfaction and effectiveness are correlated [42], it is unclear to which extent user expectations are manipulated.

Finally, Spiel et al. investigated the priming of two difficulty adaptations of a TETRIS game, where the speed of falling pieces is actively adapted based on the player's performance and eye movements. The participants were divided into two groups, where one group was informed about the difficulty adaptations while the other was not. The authors found significant differences in the brick fall speed between both groups, implying that prior framing of the participants can result in changing playing performances. While the authors used a functional adaptation throughout their conditions, the question of how a suggested but non-functional placebo affects user expectations remains open.

2.2 Usability and Assessment of Adaptive User Interfaces

How can we infer if one technology is better than another? In user studies, subjective measures for acceptance and usability such as the Technology Acceptance Model (TAM) [15, 16, 101], Unified Theory of Acceptance and Use of Technology (UTAUT) [46], or the NASA Task Load Index (NASA-TLX) [38] for the measurement of workload, are widely used metrics in the field of HCI [37]. However, such subjective measures are prone to subjective biases. They require the participant's ability to think back and evaluate the interaction with the UI, can be susceptible to socially desirable responding (e.g., see [59, 75] or confirmation-bias [76]. Objective measures such as task completion time, error rates or physiological measures (e.g. heart rate), are less affected by such biases [99]. The adaptation of UIs has become an integral part of interactive systems to satisfy the user's goals, widen the user base, and extend the system's lifespan [8]. User-centric adaptation can span from the content or the interface format and the different user input modalities. For example, one can adapt letter presentation in an E-reader to alpha oscillations; an indicator of workload in the electroencephalogram [55], reduce the number of interaction possibilities with the number of errors made by the user [45], or adapt the interface to the mood of the user [61]. Further examples include the user's context [90] and their interaction performance as well as their physiology [56] serving as implicit measurements for user interface adaptations. Many user interfaces base their adaptations on AI-methods. The availability of large-scale user data allowed recent advances in machine learning and neural networks to make robust predictions about user-related issues during interaction. This includes adaptations based on the user's perceived workload [53–55], emotions [5, 52], and interaction difficulties

Table 1. HCI studies retrieved from the literature search. We categorize the studies into **(a)** studies that employed a placebo-control, **(b)** studies that specifically investigate the existence of a placebo effect, and **(c)** studies that conclude with a need for placebo-controls.

Category	N	Studies
Placebo-Control	4	[35, 71, 97, 111]
Placebo Study	12	[18, 20, 22, 23, 28, 82, 100, 107, 110, 110]
Acknowledge Placebo Effects as a Confound	11	[1, 7, 10, 21, 47, 64, 74, 83, 86, 91, 94, 105]

due to the complex visualization of UI elements [51]. AI-based adaptive interfaces optimize usability by linking the system design to user input [43], where the overall goal of user interface adaptations is the improvement of the interaction efficiency and perceived usability [43]. However, previous placebo research [22, 80, 96, 100] indicated that the sole description of a non-functional adaptive system can positively increase user satisfaction without system functionality. Garcia et al. [33] showed that adding an animation that supports a user's mental model can increase performance ratings of a system, even if it performs objectively in the same manner or even worse than the non-animated system. Therefore, descriptions of systems can be regarded as a potential placebo, if aside from perceived effectiveness the task performance is susceptible to such manipulations [22, 32].

Although previous work successfully demonstrated placebo effects in difficulty-adapting video games, the systematic investigation of placebo effects on user expectations has received little attention in HCI research. We argue that unintentionally manipulating the user's expectancy, which is a necessary condition for placebo effects, is likely to manipulate the results of HCI studies on a larger scale. This placebo effect is potentially biasing conclusions in novel and high-expectancy areas such as interaction with AI. We address this research gap by presenting two experiments investigating the impact of user expectations and placebo effects on user experience measures and user performance.

2.3 A Placebo Effect in User-Centered Studies

We performed a systemic literature search to identify previous work concerning placebo effects in HCI. We used the regular expression (*"placebo"* AND *"placebic"*) to search for relevant publications in the ACM Digital Library that investigated placebo effects in HCI studies. This includes publications from more than 50 journals and proceedings of more than 170 conferences. Our initial query resulted in 420 publications that we further narrowed down by excluding those situated in the medical field, which briefly mentioned placebos as a medical or psychological science method. The resulting data set contained 27 publications (see Table 1).

We identified three types of HCI user studies that account for the potential influence of placebo effects (see Figure 1a). First (N = 12), some studies entertained the possibility of placebo effects within their or other studies (see Figure 1a). Second (N = 4), some studies employed placebo-controls (see Figure 1c), comparing user evaluations of a sham-treatment/system against a functional system, e.g. a non-functional brain-computer interface against a functional brain-computer interface [22]. Third, and most importantly, some studies (N = 11) compared a sham treatment against a baseline condition. These studies were designed specifically to provide evidence of a placebo effect in HCI studies. They employed system descriptions to implement a placebo, describing a function of a system that was not implemented. In line with practices in Psychology [6], and Medicine [29], we will use the term "placebo-studies" to refer to studies that employ a sham-treatment and compare it to a baseline condition, and "placebo-controlled studies", studies that compare a sham-treatment against a functional system.

Placebo studies in HCI suggest that user evaluations of technology can be biased by knowledge and expectations about a given system. Vaccaro et al. [100] demonstrated that when giving users an illusory control with buttons supposedly intelligently regulating social media content, these elements increase feelings of control and trust in

a system. Rutten and Geerts [88] demonstrated this for the evaluation of novel systems, which is a prominent case in HCI. In an experimental design, they compared two haptic touch devices. They could show that ratings of attractiveness and pleasure post-interaction can largely be attributed to the perceived novelty of a system. Therefore, evaluation can be strongly affected by descriptions of the system.

Also, more specific evaluation dimensions such as immersion and heart rate seem to be affected by knowledge or belief about a system. Denisova et al. [19, 20] told players that a game adapted the difficulty to their performance, even though no adaptation took place. The players receiving this sham system description judged the game to be more immersive. In a similar experiment, Duarte et al. [22, 23] found that players' heart rates increased in response to sham game elements that allegedly empowered the game character. Paepcke and Takayama [80] implemented a similar placebo manipulation in a study of two robot interfaces. They described an interface with differing levels of functionality; one with a lot of features, the other with only a few. However, the participants were using the same interface. The participants preferred interaction with the feature-rich robot as compared to the robot with only minor features. In sum, these studies indicate that descriptions of a system can be considered placebo in user-centered evaluations of usability. As technology is typically evaluated in terms of usability and interaction efficiency, we operationalize a placebo effect in terms of different system descriptions that affect our measured metrics. Changes in metrics could be due to an interaction with the alleged AI system, including the percentage of correct answers or subjective ratings of the user's performance after interaction with the AI.

3 EVALUATING PLACEBO EFFECTS OF AI IN USER STUDIES

We conducted two experiments to evaluate the placebo effect in Human-AI interfaces. Previous research shows that placebos can be induced by social learning [50] (e.g. observing someone else improving), classical conditioning (e.g. associating an inert pill with relaxation) [30] or verbal information (e.g. expert instructions) [95]. Motivated by the literature survey, we sought to induce a placebo effect by manipulating system descriptions. Both experiments were conducted remotely and were started by launching a web app in a browser. Participants were asked to solve word puzzles with varying difficulties. In the first experiment, we explored how informing participants about experiencing an alleged *Difficulty Adaptivity* can change the user's expectations, subjective performance ratings and objective task performance. In the second experiment, we tested the placebo effect by comparing subjective performance ratings and objective performance in a control condition and a *Physiology-based Adaptation* condition within-subjects. Regardless of the assigned adaptivity, the task difficulty remained constant in all studies, i.e. no changes to the task difficulty took place at any time. Understanding whether expert instructions can prompt placebo effects would enable researchers to control or correct for their influence in future evaluations.

Our studies investigated the following research question and hypotheses:

RQ: Do system descriptions confound the evaluation of adaptive systems?

We investigated the following hypotheses to answer this research question:

H1: User expectations can be manipulated by a description of an adaptive AI system.
H2: User performance and performance ratings can be manipulated by a description of an adaptive system.
H3: Workload decreases when participants believe they are being supported by an adaptive AI system.

3.1 General Study Approach

We performed two experiments to investigate the hypotheses and answer the research questions. In Experiment I, we examined the possibility of a placebo effect using a between-subject design: Participants were informed about all possible adaptation scenarios (e.g., *No Adaptivity* vs. *Physiology-based Adaptation*) beforehand and were then told which of these conditions they would be assigned to. The between-subject design was to ensure that probing the functionality of one system did not interfere with evaluating the other. In Experiment II we ran a within-subject experiment. This would allow users to experience a non-functional assistance system against an

Table 2. Experimental conditions and their assigned properties. *No Adaptivity* and *Error-based Adaptation* convey a transparent functionality while maintaining a low complexity. In contrast, the *Physiology-based Adaptation* provides a low transparency and high complexity.

	Transparency	**Complexity**	**Interactivity**
No Adaptivity	High	Low	Low
Error-based Adaptation	High	Low	High
Physiology-based Adaptation	Low	High	High

allegedly functional system by each participant, hence increasing statistical power and experimental control. In detail, participants could compare their performance for the *Physiology-based Adaptation* and the *No Adaptivity*. Here, we could investigate if a placebo effect of AI emerged even when the baseline and placebo-control conditions were known.

4 EXPERIMENT I: EXPLORING THE PLACEBO EFFECT OF AI

This experiment presents an experimental between-subject design to explore the possibility of a placebo effect in Human-AI interfaces. We describe our experimental design below.

4.1 Experimental Design and System Description

The experiment was conducted remotely using a web app. Participants were instructed that they had been selected for one out of three *Adaptivities*, and that their task was to solve word puzzles of varying task difficulty compromising our manipulation of the system description. They were advised that the difficulty of the word puzzles would be adapted according to the assigned *Adaptivity*. Each participant processed the same word puzzles in a randomized order without any difficulty adjustments. The participants were informed about the presence of all *Adaptivities* before the experiment. We have intentionally decided to inform the participants about all available conditions before asking them for their anticipated task performance and expectations using the assigned Adaptivity. Hence, the participants can compare their expectations, provided by the description of the assigned *Adaptivity*, against the other two *Adaptivities*.

The participants were informed about the presence of all conditions before starting the experiment, as is standard practice in medical studies. However, we also disclosed the assigned adaptivity to the participants since we were investigating the AI placebo effect of knowing an alleged working treatment (i.e., *Error-based Adaptation* or *Physiology-based Adaptation*), compared to the baseline (*No Adaptivity*). The system descriptions were shown to all participants before assigning them randomly to one. We summarize the system descriptions below. Table 2 shows an overview of the conditions and their assigned properties manipulating the participants' expectations, i.e. a non-blinded between-subjects design with the three following groups:

No Adaptivity: The participants were told that no assistant was selected to adjust the difficulty of the word puzzles. The task difficulty was randomized and was not influenced by the user's performance or physiological responses. Therefore, no changes in task difficulty took place. This condition served as a control.

Error-based Adaptation: Adapting the task difficulty according to the task performance is an established paradigm [31, 93] enabling the user to self-validate the functionality of the task difficulty adaptation. Participants were told the task difficulty would be adjusted according to the number of errors they made, so that correct answers would increase the task difficulty and a streak of wrong answers would trigger decreased difficulty. This condition presents a control condition concerning validation of system functionality by the user, i.e., the
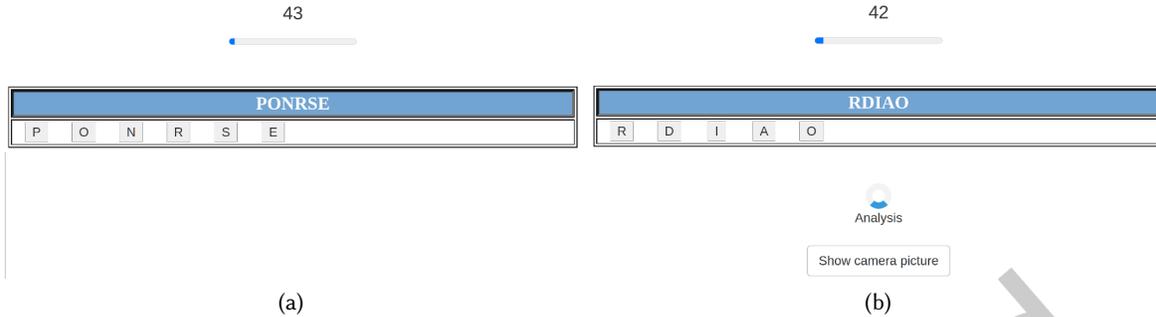
Fig. 2. Participants were randomly assigned to one of the three *Adaptivities*. Each word puzzle had a time limit of 45 seconds, upon which the experiment continued with the next word puzzle. **(a):** Participants either learned that the task difficulty was not adapted at all or was adapted according to the number of correct and wrong answers in a staircase model. The correct word was "PERSON". **(b):** We explained that the difficulty was adapted using an AI, which uses the camera image to deduce workload and stress-levels. An animation pretended a constant analysis where the camera feed could be viewed. The correct word was "RADIO".

user could check whether an error resulted in the algorithm choosing an easier item in the subsequent trial. In contrast to the *Physiology-based Adaptation*, the *Error-based Adaptation* allows participants to self-validate their results, i.e., the algorithm that reflects the item choice of the system is specified.

Physiology-based Adaptation: Participants were informed about the presence of a physiology-based adaptation of an AI changing the word puzzle difficulty by analyzing the heart rate [62, 68] and emotions via facial expressions [25] (i.e., an AI-based adaptation). A scientifically valid explanation was provided in the system descriptions to the participants to convince them of the efficacy of the physiologically-based approach. The participants were told that changes in heart rate [67, 81] and perceived emotions [84] are correlated to cognitive demand and stress-levels that can be captured using a webcam, and that an AI (i.e., neural network) aggregates those two metrics to calculate the currently perceived cognitive workload and stress level. The difficulty of the word puzzles was then supposedly adapted according to these measures to optimize user performance. The participant's webcam was activated, and the camera picture could be viewed and validated by the participant during the experiment to ensure that its functionality was being utilized. No data from the camera was recorded or transmitted. Participants were also assured that the processing of the video stream took place on the device itself. The exact algorithm that reflected the word puzzle difficulty was not specified.

Both the *Physiology-based Adaptation* and the *Error-based Adaptation* are AI interventions. While the *Physiology-based Adaptation* is described as a complex non-linear and non-traceable algorithm, the *Error-based Adaptation* is traceable and verifiable by the participant. Previous research showed that the mental model of a system is crucial for manipulating performance ratings [33]. Therefore, we compare both descriptions of theoretically simple verifiable algorithms (i.e., *Error-based Adaptation*) to a more complex but supposedly more intelligent system (i.e., *Physiology-based Adaptation*).

4.1.1 Task Description and Apparatus. The experiment was remotely conducted via a web app in the participant's web browser. Each interaction took place in the browser, including the experimental instructions and the system descriptions. The participant's demographics, consent, task performance, and self-report measures were logged exclusively on the server. The participants were then randomly selected for one of the three aforementioned conditions after explaining each of them.

The experiment presented participants with word puzzles to solve. Each puzzle consisted of a randomized sequence of characters, which they had to select in the correct order to spell an English word. Their selection was incrementally displayed in a text field above this random sequence. Upon selection, characters were removed and thus could not be selected again. After choosing all available characters the puzzle ended, and the next word puzzle was presented. A visible time limit of 45 seconds was set for each trial. The participant's responses were automatically recorded when the last character was selected or after the 45 seconds, whichever came first. The *Adaptivity* condition was displayed throughout the experiment. This word puzzle task allowed us to modulate the task difficulty noticeably for participants (i.e., through the number of presented characters) while not revealing the correct answer immediately. Figure 2 depicts an example of the word puzzles for each condition.

Overall, participants solved 20 different word puzzles during the experiment. All participants received the same word puzzles, presented in random word order and a randomized arrangement of characters. Hence, the participants solved 20 word puzzles in every condition. The word pool in Experiment I consisted of the same 20 words. Greater task difficulty was defined by increased word lengths, where the word length was normally distributed (i.e., between three and eight characters). Words were selected from the Oxford 5000[1] list, which contains the essential English basic words. The rules of the word puzzle were well-defined and had an unambiguous solution. The game rules were explained to the participants while the game complexity was varied through the number of displayed characters (i.e., higher number of characters lead to a higher task complexity) and the familiarity rating of the words. Hence, the game difficulty could be quickly varied to suggest difficulty adaptations to the player.

4.1.2 Measures. After informing the participants about the *Adaptivity* they had been assigned to, we asked them about their expected performance of the *Adaptivity* on a seven-point Likert scale: "You are in the group where the word puzzles are adjusted by your task performance. Do you think you will do better compared to the other groups?"[2]. This quantified the effect of our system description manipulation [17]. When solving the word puzzles, we measured the required time in seconds and the number of errors for each answer. An error was logged whenever a wrong word was chosen by selecting the last remaining character, or when the time limit of 45 seconds was up. We employed the raw NASA-TLX [37, 38] questionnaire after the experiment to measure the perceived workload.

4.1.3 Procedure. The course of the study was explained to the participants by visiting a web page that described the evaluation of an intelligent online assistant. Afterwards, participants provided consent in accordance with the declaration of Helsinki. Then we explained the three adaptivities. This was followed by demographic questionnaires, including the participant's self-estimated technical competence and belief in improved performance over the other two adaptivities, which was probed on a seven-point Likert scale, respectively. Next, a validation of the participant's understanding of the placebo stories was conducted. Participants were asked to fill out three multiple-choice questions that tested their understanding of the intention of the study, modulation of task difficulty, and the available adaptivities (see Appendix). This information was later used to exclude participants who misconceived the study details. Before the quiz, participants were randomly assigned to one of the three adaptivities. Following that, they were asked to solve the 20 word puzzles and to fill in a NASA-TLX questionnaire [37, 38]. The experiment took approximately 15 minutes. The participants were debriefed regarding the deception after the experiment. They could then choose if they wanted to withdraw their data or still provide them for further processing. The experiment procedure is depicted in Figure 3. The German Society for Psychology and the local institutional ethics board provided ethical approval for the study[3].

---

[1]www.oxfordlearnersdictionaries.com/wordlists/oxford3000-5000 - last access 2022-04-10

[2]1 = I will perform much worse; 7 = I will perform much better.

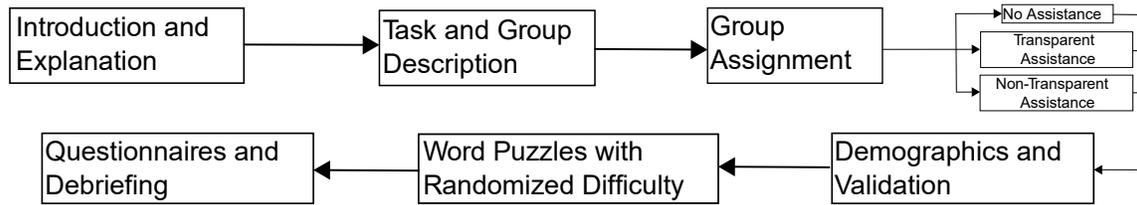[3]Ethical assessment ID: EK-MIS-2020-023

Fig. 3. The study procedure of Experiment I.

4.2 Participants

We used Prolific[4] to recruit participants for our study. In Experiment I, we set the filter criteria to native speakers from the US and the UK. We used the same filter criteria for Experiment II as in Experiment I in addition to excluding participants from Experiment I. Participants were compensated with nine Pounds per hour. We removed participants who did not pass the validation questions or who completed the word puzzle in a very short time. Subsequently, participants providing invalid data were removed from the analysis and did not receive compensation. Three hundred forty nine participants completed our experiment successfully. The data was cleaned before analysis. Participants who did not answer all three validation questions correctly, achieved less than 10% of correct answers in the word puzzles (i.e., 2 answers or fewer), took less than 100 seconds to complete all puzzles, or did not answer the word puzzles or questionnaires completely were excluded from the data analysis. Participants were compensated with three British pounds for their participation. Of the 369 participants, 75 (21.48%) did not carefully read the instructions as they could not answer all three questions regarding the system descriptions. Further, 16 participants (4.58%) did not comply with our instructions and solved less than 2 of the 20 word puzzles, three of whom (0.86%) did run through in a reasonable time frame, but they were therefore still excluded from our statistical analysis. This left 255 (69.11%) participants after preprocessing the data. One hundred sixty six participants were female, 80 were male, four identified with another gender, and one participant preferred not to disclose their gender. On average, participants were of medium age (M = 32.13; SD = 12.41), and living in English-speaking countries. Participants rated their technical competency with novel technologies on a seven-point Likert scale, all reporting a medium to good proficiency with technological systems (M = 4.61; SD = 1.30). Eighty four participants were briefed with *No Adaptivity*, 93 received the *Error-based Adaptation* and 78 obtained the *Physiology-based Adaptation* narrative[5].

4.3 Analysis

In the following, we analyze differences between the three groups regarding their subjective and objective performance indices. We will test for the effects of our manipulation by comparing each adaptive condition to the *No Adaptivity* condition Mann-Whitney U-tests or Wilcoxon-rank test for paired samples. We will test user expectations before and after interaction with the system in a rank-aligned repeated-measures analysis of variance (ANOVA) [49]. Follow-up analysis is based on Bonferroni-corrected ART-C contrasts [26] Correlations for Likert-type scales (ordinal) are estimated using Spearman-rank correlations. The *No Adaptivity* always

---

compromises the baseline for all measurements. NULL-hypotheses relevant for our main research questions are evaluated using Bayes factors computed using the BayesFactor package [73] or JASP [106][6].

4.4 Subjective Evaluation

We analyzed the self-assessed subjective task performance before and after the word puzzle quiz, contrasting each assistance condition and the control-condition as a between-subjects factor. If system descriptions change expectations and the evaluation of effectiveness, we would expect mean differences in performance expectancy in favor of the *Physiology-based Adaptation* and *Error-based Adaptation* conditions as compared to the *No Adaptivity*-condition to carry to performance ratings after the interaction.

In accordance with our hypothesis, the narrative of the adaptivity did affect the users' subjective estimates of task performance prior to solving the word puzzles for the *Error-based Adaptation*. Participants in the *Error-based Adaptation*, U = 6434.50, p = .001, Z = -3.24, r = 0.24, expected their performance to be superior to the *No Adaptivity* modality (i.e. the control group). Figure 4 shows this change in performance expectations. This was not significant for the *Physiology-based Adaptation*, U = 6364.50, p = .086, Z =-1.72, r = 0.135. [7] Note that this pattern was mirrored in performance ratings after interaction with the system. In the *Error-based Adaptation*, U = 6714.50, p = .022, Z = -2.29, r = 0.172 but not in the *Physiology-based Adaptation*, U = 6479.50, p = .673, Z = 0.42, r = 0.03. [8], when compared to *No Adaptivity* (see Figure 4b).

We computed a repeated-measures rank-aligned ANOVA (between = *Adaptivities*, within= *Time point (before vs. after UI interaction)* using the ArTool that revealed a significant effect within the *Adaptivities*, $F(2, 252)$ = 5.36, p < .001. We also found a significant main effect of *Time point (before vs. after UI interaction)*, $F(1, 252)$ = 179.50, p < .001. Participants judged their performance to be better before solving the word puzzles (M = 4.12; SD = 1.04) as compared to after completing the word puzzles (M = 2.93; SD = 1.37). The interaction of the factors in the ANOVA was not significant, $F(2, 252)$ = 0.54, p = .584. [9] Bonferroni-adjusted post-hoc contrasts on ranks across *Adaptivities* could show that our manipulation resulted in an expectation of superior task performance, averaged across Time Point only for the Error-based Adaptation and the control-condition, $t(252)$ = -3.16, p = .005 but no difference in ranks between the Physiology-based Adaptation and the control-condition, $t(252)$ = 0.85, p >.99, or for the comparison of Physiology-based Adaptation and Error-based Adaptation, $t(252)$ = -2.23, p =.080. We can conclude that performance ratings before and after interaction differed for the *Error-based Adaptation* from ratings the *No Adaptivity*-group but not for the Physiology-based Adaptation.

Both assistants were believed to be specifically adaptive after interaction. Participants responded to the question "The word puzzles have adapted well to the task difficulty", rated on a seven-point Likert scale (1 = Do not agree at all; 7 = Absolutely agree), with a relatively more agreement as compared to the *No Adaptivity* condition, (M = 3.54; SD = 1.43), *Error-based Adaptation*: (M = 4.30; SD = 1.18), U = 6253.00, p < .001, Z = -3.69, r = 0.27, and *Physiology-based Adaptation*: (M = 4.00; SD = 1.37), U = 7004.00, p = .026, Z = -2.20, r = 0.17. Therefore, making perceived adaptation in the *Error-based Adaptation*, where participants could validate functionality similar to the

---

[6]Note that we have tried different weakly informative priors. However, none of them have changed the results substantially, so we stuck with the default priors of the BayesFactor-Package.

[7]Here, using the Bayesian equivalent of a Mann-Whitney U-Test, [106], we can show that that this is not due to the NULL-Hypothesis (no difference between ranks) being true, BF01 = 1.20, but rather due to insufficient amount of information.

[8]The Null-hypothesis, is about 5 times more likely than the alternative hypothesis, BF01=5.989

[9]To verify if it is justified to conclude that expectation readjustment is uniform across *Adaptivity* levels, which amounts to rejecting the NULL-hypothesis, we carried out an additional Bayesian analysis. We performed a parametric repeated-measures Bayesian ANOVA [106] on performance ratings for each time-point. Note that this parametric analysis could be considered more liberal. The Bayes factor comparing the model with and without interaction effect suggested that the data were 14.78 : 1 in favor of the simple effect model, making it 15 times more likely to be true as compared to the alternative hypothesis of differing means for each *Adaptivity* over time given the data. Thus, the data are in agreement with user expectation decrease being uniform.
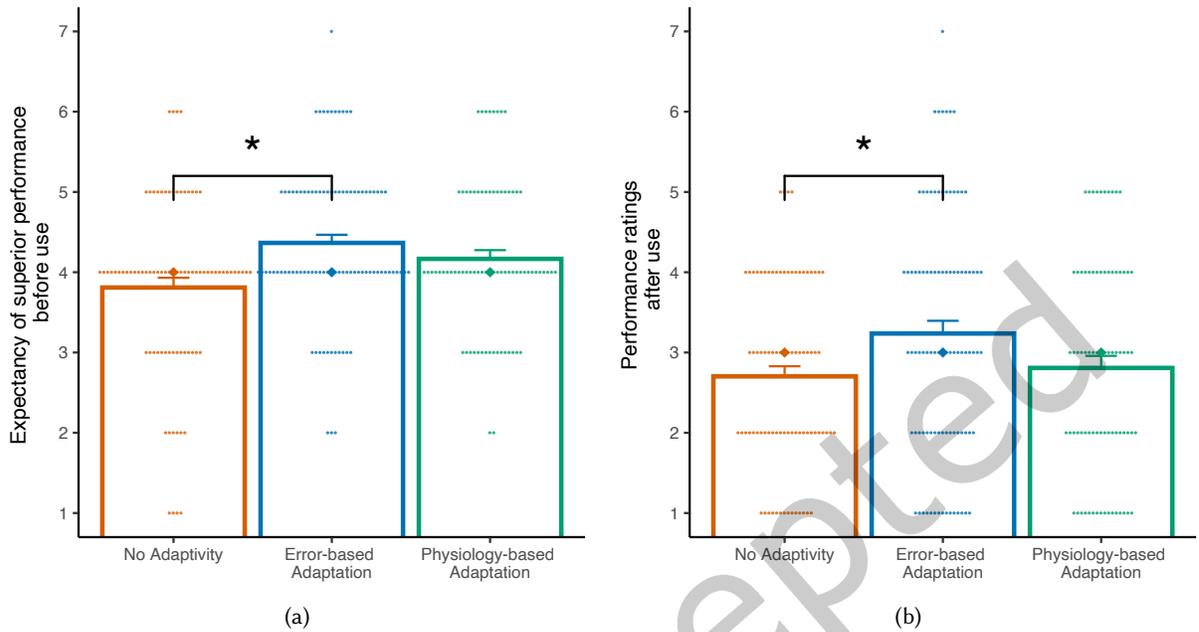
Fig. 4. **(a):** Averaged measured expectations regarding the anticipated task performance of the assigned *Adaptivity* **before** the experiment. **(b):** Averaged measured expectations regarding the perceived performance **after** the experiment. The error bars depict the standard error of the mean. Brackets indicate significant differences. Dots are individual data points. The rhombus indicates the median value.

Physiology-based Adaptation condition, we follow that it is highly unlikely that subjects' validation of system functionality had any impact on performance expectations and thus effectiveness.

4.5 Number of Errors

Task performance was operationalized as the number of correct responses participants made during the word puzzle task (i.e., percent of correct answers). Participants solved the tasks with an average accuracy of 69.90% (SD = 14.88%) and agreed on a 7-point Likert scale (M = 3.53; SD = 1.19) that the word puzzles were easy to solve. Thus, objective and subjective difficulty of the word puzzles were considered easy to medium, allowing for an ideal level of variation. We further investigated if manipulation of system description changed task performance. None of the conditions differed significantly (*Physiology-based Adaptation*: M = 67.63% SD = 14.61%, $t(159.86)$ = 1.28, p = .203, d = 0.20; *Error-based Adaptation*: M = 71.14% SD = 14.68% , $t(171.54)$ = 0.23, p = .821, d = 0.03) as compared to the control-condition (M = 70.63%, SD = 15.28%).

The expectancy of task performance (before interaction) was substantially correlated (spearman-correlations) with the relative number of correct responses during interaction, r = .25, p < .001. This correlation was not uniform across groups, *Physiology-based Adaptation*: r = .40, p < .001, p < .001 , *Error-based Adaptation*: r = .27, p = .008, control-condition: r = .12, p = .289. Thus, we concluded that if performance expectations can influence actual task performance, then this is most likely for the *Physiology-based Adaptation*.We found a similar pattern for the performance ratings after the interaction, r = .40, p < .001, which were largely uniform across *Adaptivities*,

Table 3. Means (*M*) and standard deviations (*SD*) of technology acceptance scales as a function of groups.

	Error-based Adaptation	Physiology-based Adaptation				
	M (SD)	M (SD)	df	t	p	d
Performance Expectancy	14.80 (4.96)	12.17 (6.08)	148.41	3.06	**.003**	0.48
Effort Expectancy	16.62 (4.92)	15.14 (6.57)	140.48	1.64	.102	0.26
Hedonic Motivation	11.42 (4.15)	10.40 (4.74)	154.46	1.48	.140	0.23
Personal Innovativeness	13.35 (4.29)	12.47 (4.68)	158.18	1.27	.205	0.20
Novelty	13.02 (3.54)	12.47 (5.00)	135.21	0.81	.419	0.13
Attitude	12.97 (3.51)	11.63 (4.89)	136.45	2.02	**.045**	0.32
Overall Reward	12.23 (3.80)	11.00 (4.90)	143.66	1.80	.074	0.28
Behavioral Intention	9.96 (3.93)	8.79 (4.62)	151.91	1.75	.082	0.27
Overall Risk	10.04 (4.10)	10.33 (4.55)	156.60	-0.43	.665	-0.07

.39 > r < .44. The higher they rated their own performance, the higher their individual percentage was correct. Therefore, self-assessment of performance was valid across condition.

4.6 Workload

There was no effect on workload measured by the raw NASA-TLX for the placebo conditions (*Physiology-based Adaptation*: M = 65.91; SD = 13.68), $t(156.70)$ = -0.21, p = .831, d = -0.03; *Error-based Adaptation*: M= 64.26; SD = 14.91), $t(165.74)$ = 0.47, p = .640, d = 0.07, control; M = 65.39; SD = 17.07) . This is likely because subjective performance expectations (before UI use) were unrelated, *r(spearman)* = -.01, p = .818. However, performance ratings (after UI use) were negatively related to the perceived workload, *r(spearman)* = -.24, p < .001. Task performance (percent correct) was also strongly negatively related to workload, $r(253)$ = -.35, p < .001. Participants who reported low workload performed significantly better. Note that this relation was largely uniform across groups, -.30 > r < -.40,' and thus did not warrant a more fine grained analysis.

4.7 Specificity of Perceived Adaptiveness

We calculated a sum score for the perceived adaptiveness regarding the individual stress-level and adaptiveness based on prior performance as presented in previous research [18] (see Appendix). The score could range from five to 35. High scores here indicate a high level of perceived adaptiveness based on either the physiological stress-level or performance. Adaptiveness based on task performance did differ between the *No Adaptivity* (M = 17.36; SD = 7.49) and *Error-based Adaptation* (M = 23.52; SD = 6.23), $t(162.13)$ = -5.91, p < .001, d = -0.90. This difference was also significant when comparing the *No Adaptivity* to the *Physiology-based Adaptation* (M = 21.27; SD = 6.71), $t(159.81)$ = -3.51, p < .001, d = -0.55. However, still the *Error-based Adaptation* was rated to be more adaptive than the *Physiology-based Adaptation*, $t(159.06)$ = -2.25, p = .026, d = -0.35. A different pattern of results was found when comparing *No Adaptivity* (M = 15.13; SD = 6.44) with the adaptive conditions on the scale measuring perceived adaptiveness based on stress-levels, *Physiology-based Adaptation*: (M = 19.09; SD = 6.06), $t(159.97)$ = -4.03, p < .001, d = -0.63, *Error-based Adaptation* (M = 16.57; SD = 5.74), $t(167.13)$ = 1.56, p = .120, d = 0.24. Mean adaptiveness based on stress-levels differed significantly between the *Physiology-based Adaptation* and the *Error-based Adaptation* $t(160.41)$ = -2.78, p = .006, d = -0.43. Therefore, both adaptivities were perceived to adapt to performance, however only the *Physiology-based Adaptation* was perceived to adapt to the user's stress. We can therefore conclude that the illusion of adaptiveness is specific to the conditions even after system evaluation, and that users did not validate the adaptive system.
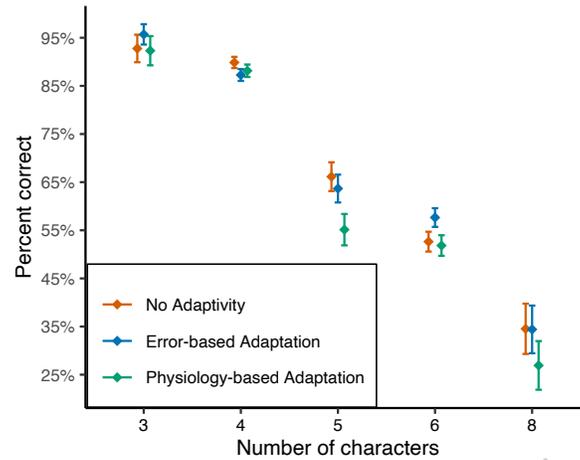
Fig. 5. Average percentage correct for the assigned *Adaptivity*. Error bars denote +/- 1 standard error of the mean.

4.8 Technology Acceptance

For both adaptive conditions, we measured a range of scales regarding technology acceptance on a seven-point Likert scale. All of the scales compromised sum scores. The scales' performance expectancy (e.g., "I find the assistant useful in my daily life"), effort expectancy (e.g., "I find the assistant easy to use") and hedonic motivation (e.g., "Using the assistant is fun") were adapted from the UTAUT [103]. Each scale had four items rated from 1 (Do not agree at all) to 7 (Absolutely Agree). Therefore sum-scores could have a range of four to 28. Personal Innovativeness (e.g., "I like to experiment with new information technologies"), Novelty (e.g., "Using the assistant is new and refreshing"), Attitude (e.g., "Overall, it is wise to use the assistant"), Overall reward (e.g., "Using the assistant will be rewarding"), behavioral intention (e.g., "I intend to use the assistant to adapt the task difficulty") and overall risk (e.g., "Using the assistant would be risky") were adapted from Wells et al. [108]. The scales had three items, each rated from 1 (Do not agree at all) to 7 (Absolutely Agree). Therefore, sum scores could have a range of three to 21.

Most of the scales differentiated between the conditions (see Table 3). Only performance expectancy was slightly increased in the *Error-based Adaptation* condition compared to the *Physiology-based Adaptation* which largely resembles the decline in performance expectations (see Figure 4a). The attitude scale followed a similar trend. We therefore chose not to include the UTAUT in our second study.

4.9 Length of Word Puzzles

We found a potential problem area regarding word length for participants' performance within Experiment I when exploring our data set; namely, that variation between groups is small for easy, three character words and for more difficult words of eight characters, but may differ more when encountering items of medium difficulty of five characters, see Figure 5. As this variation could have decreased our ability to detect effects, we chose to use a validated set of word puzzles in the English language with five characters for Experiment II, where we conducted a within-subjects study.
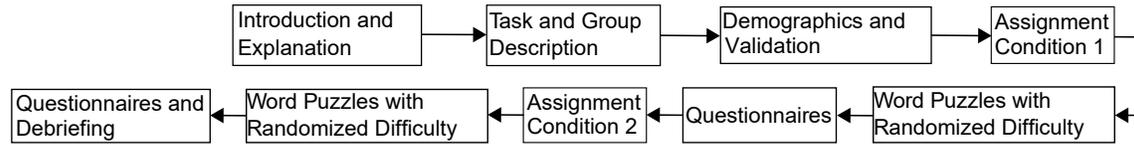
Fig. 6. Study procedure of Experiment II.

4.10 Discussion

Our results show that the description of an adaptivity changes the perceived system efficiency of the user. Task performance expectations were increased before use when comparing the *Error-based Adaptation* and *No Adaptivity*-condition (H1). This difference was not found for the *Physiology-based Adaptation*-condition. In consequence the placebo effect estimates of higher performance after interaction only surfaced for the *Error-based Adaptation*.

Participants perceived the adaptiveness to be specific to their respective groups. Thus, participants in the *Physiology-based Adaptation* reported that the word puzzles were adapted to their perceived stress-level, and participants in the group that believed they were receiving *Error-based Adaptation* were strongly convinced that the word puzzles were adapted to fit their performance. We conclude that the possibility of validating the adaptive algorithm in the *Error-based Adaptation*-condition did not seem to hinder observing an effect on performance ratings.

Still, belief in the superiority of an adaptive system differed between individuals, and therefore, may need to be systematically controlled to obtain a measurable behavioral effect. No effect was found for the Physiology-based Adaptation condition, the number of errors (H2) and perceived workload (H3) between the groups, likely due to the substantial variance in performance expectation across subjects and the associated low statistical power. This variance, however, was predictive of the number of errors in the task, especially for the *Physiology-based Adaptation*-condition. Therefore, we chose to replicate and extend our findings in a within-subjects experimental study to control for the individual perception and variance of the *Adaptivities*.

5 EXPERIMENT II: A PLACEBO EFFECT WHEN COMPARING INTERFACES

We revealed that an *Error-based Adaptation* system adaptivity produces placebo effects that impact subjective performance expectations and performance ratings. After interaction with the system, participants still believed in its adaptive function and could distinguish between *Physiology-based Adaptation* and the *Error-based Adaptation* condition. We did not find an effect on objective measures regarding task performances, such as percentage of correct answers, possibly due to varying difficulty levels in the puzzles. However, there was a strong correlation of expectancy and task performance for the *Physiology-based Adaptation*-condition; thus, if user expectations can be manipulated more consistently, we expect a placebo effect of the *Physiology-based Adaptation* on objective task performance and performance ratings. Now that we have established the possible size of a placebo effect in a user study with the *Error-based Adaptation* condition, we will compare the *Physiology-based Adaptation* adaptive system to a control condition in a within-subject design. Hence, we compare the *No Adaptivity* with the *Physiology-based Adaptation*. We removed the *Error-based Adaptation* due to its transparency and simplicity. The *Error-based Adaptation* can be verified regarding its correctness, hence mitigating a potential placebo effect. Thus, every participant encountered both the *No Adaptivity* and the *Physiology-based Adaptation* condition in a counterbalanced order. We recruited 100 participants who had not participated in Experiment I, where half of the participants (N = 50) started with the *Physiology-based Adaptation* followed by the *No Adaptivity* condition. The other half started with *No Adaptivity* followed by the *Physiology-based Adaptation*.

5.1 Method

We used a similar web app as described in Experiment I. The web app was modified to implement a within-subjects design where participants were confronted with two conditions, see Figure 6. The participants were instructed with a cover story to solve word puzzles using the *No Adaptivity* and *Physiology-based Adaptation* condition as described in Experiment I. In other words, participants solved word puzzles using each *Adaptivity* in a counterbalanced study design.

Each condition contained 40 word puzzles with words that were selected from Gilhooly and Johnson [36]. Participants were asked to solve word puzzles of varying difficulty with the use of the two different adaptivities, while items were presented randomly throughout both conditions. The participants solved 40 words per adaptivity, resulting in 80 word puzzles for the experiment. The words and their character arrangement were randomized for each condition. The description of the two adaptivities, that is *No Adaptivity* and *Physiology-based Adaptation* remained the same as in Experiment I.

We added a "calibration phase" to the procedure of Experiment II to familiarize participants with the word puzzles. Before starting with the assigned adaptivity condition, participants were instructed to solve five word puzzles to get used to the task. Before starting with the *Physiology-based Adaptation* condition, participants were instructed to solve five word puzzles to calibrate an artificial intelligence driving the assistance system. We added a numeric measure of performance expectancy. Before each quiz, we asked participants how many word puzzles they were estimating they could solve correctly. Here, participants could indicate if they expected between zero to 40 correct answers in the upcoming quiz. This provided us with an estimate of the participant's expectancy regarding the anticipated efficiency while solving the task. After each adaptivity, participants filled in a NASA-TLX questionnaire. As we had established acceptance and usability concerning our system descriptions in Experiment I, we did not collect data on the UTAUT or specificity of adaptiveness in Experiment II. Participants were debriefed after the experiment and could choose if they wanted to withdraw their data or still provide them for processing. The overall experiment took approximately 40 minutes. The experiment procedure is depicted in Figure 6.

5.2 Results

5.2.1 Participants. We recruited 100 participants[10] through Prolific using the same selection parameters as in Experiment I, adding that they must not have been participants in Experiment I to the recruitment query. Data cleaning was done resembling the previous study: 19% (19 of 100) participants who did not answer all three validation questions correctly were removed. We also removed 10 instances where participants completed less than 10% (4 of 40 puzzles) one participant who only completed half the study. No suspiciously short experimental runs were found in this study. This procedure left 75 participants in our sample to analyze. 40 participants reported to be female, 34 reported participants to be of male gender and one participant did not disclose their gender.

5.2.2 Subjective Evaluation. We computed a rank-aligned repeated measures ANOVA to analyze user performance expectations prior and post interaction (within-subjects factor) with the adaptivity, (within-subjects factor). There was a significant effect of the adaptivity, $F(1, 222) = 74.70$, $p < .001$. Participants believed that their performance would be superior when assigned the *Physiology-based Adaptation* ($M = 4.31$, $SD = 1.02$) placebo condition as compared to the control-condition, ($M = 2.93$, $SD = 1.26$). There was also a significant effect of *Time point* (before vs. after UI interaction), $F(1, 222) = 67.21$, $p < .001$, indicating that after interaction expectations were readjusted. Interaction with the system diminished expectations at an average of about 1.5 points when comparing

---

[10]Sample size was determined analytically of paired *t*-test on percentage of correct scores ($SDdiff$ = .1; α = .05; N = 100, 1-β=80%), which yielded a minimal detectable effect size of μ = .028 thus rather small effects were detectable with our experimental design and sample given the statistical model.
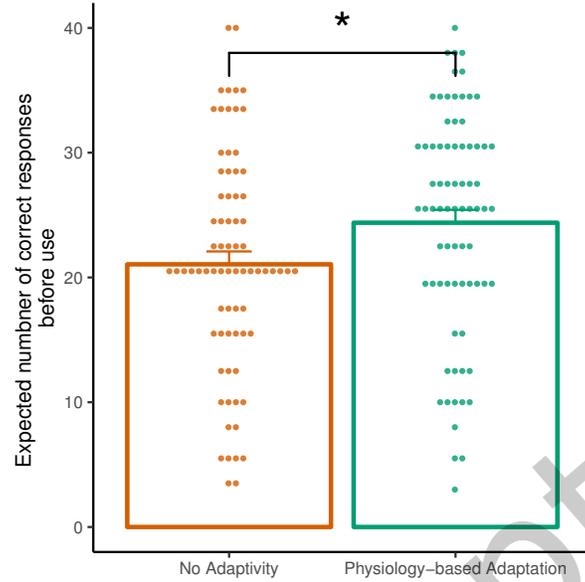
Fig. 7. Averaged expected number of correct responses as a function of the *Adaptivity*. Error bars denote +/- 1 standard error of the mean. Blue dots are individual data points.

expectations before (*M* =4.29, *SD* = 0.95) and after system exploration (*M* = 2.95, *SD* = 1.35). The interaction effect of adaptivity and time was again non-significant, $F(1, 222) = 0.04$, $p = .837$.

Still, we compared group means prior and post interaction and both were distinguishable, see Figure 8. Note that this lack of an interaction-effect [11] implies that the readjustment of expectation was consistent across the adaptivity and thus the placebo effect persisted after interaction with the sham system.

In line with subjective expectations on the Likert scale ratings, numeric estimates of performance also increased in response to our placebo-manipulation $t(74) = -2.73$, $p = .008$, $d = -0.32$. Participants expected to solve about three more word puzzles when instructed that they would be receiving *Physiology-based Adaptation* (*M* =24.39, *SD* = 8.93) as compared to *No Adaptivity* (*M* =21.05, *SD* = 8.90, see Figure 7).

5.2.3 Number of Errors. We found no difference in task performance between groups, $t(74) = -0.15$, $p = .882$, $d = -0.02$[12]. Both the control-group (*M* = 60.77 %, *SD* = 20.19) and the *Physiology-based Adaptation* (*M* = 61.20%, *SD* = 19.63) group did not differ in terms of solving of the word puzzles. Again, we investigated whether individual differences in subjective performance expectations could explain differences in performance between condition, computing the difference in performance expectations between conditions and the difference between conditions in objective performance, i.e. percent correct. In line with the previous study, we found a correlation of performance (percent correct) and expectation, $r(73)=. 28$, $p=.014$. The larger the difference in expectations in

---

[11]To verify if it is justified to conclude that expectation readjustment is uniform across the adaptivity, which amounts to rejecting the NULL-hypothesis, we carried out an additional parametric Bayesian analysis. We performed a repeated-measures Bayesian ANOVA [106] on user expectations for each time-point. The Bayes factor comparing the model with and without interaction effect suggests that the data were 4.99 : 1 in favor of the simple effect model, making it five times more likely to be true as compared to the alternative hypothesis of differing means for each adaptivity over time given the data. Thus, the data are in agreement with user expectation decrease being uniform.

[12]Using the Bayesian equivalent of a *t*-test [57], we estimate the the NULL-model with no differences between conditions is 7.7 times more likely, given the data, than the alternative model assuming differences in means
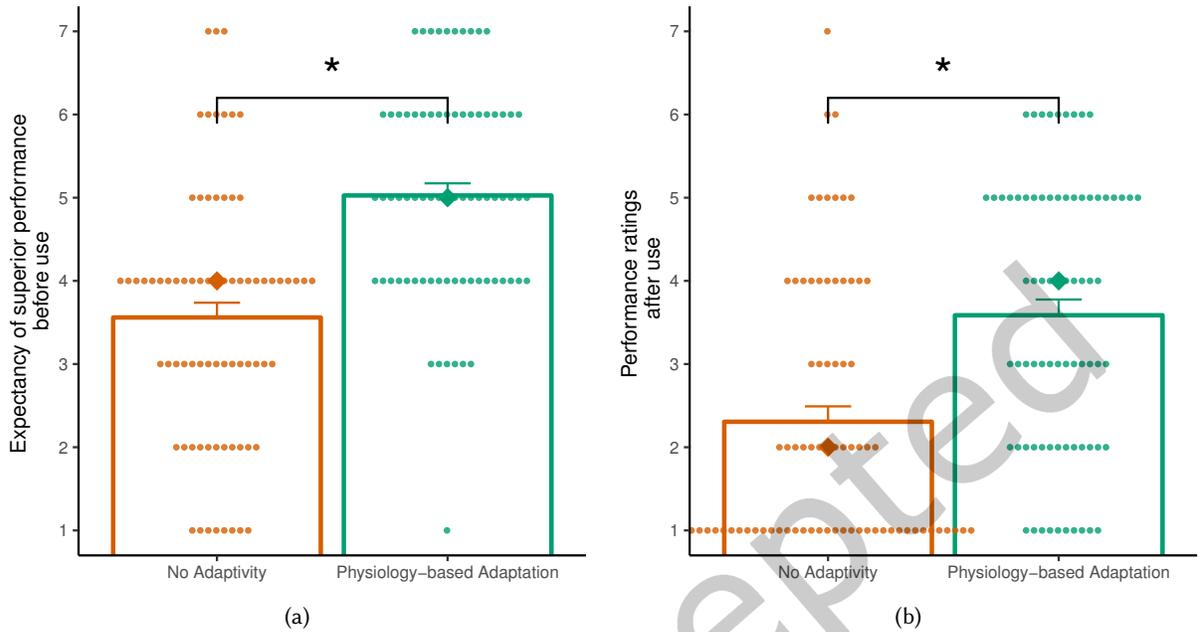
Fig. 8. **(a):** Averaged measured user expectations regarding the anticipated task performance of the assigned *Adaptivity* **before** the experiment, Z = -5.59, p < .001, r = 0.64. **(b):** Averaged measured user expectations regarding the perceived usefulness of the assigned *Adaptivity* **after** the experiment, Z = -3.94, p < .001, r = 0.45. The error bars depict the standard error of the mean. Brackets indicate significant differences.

favor of the *Physiology-based Adaptation* prior to the interaction the larger the difference in performance in favor of the *Physiology-based Adaptation*.

5.2.4 Workload. Again, we found no significant effect on workload as operationalized by the NASA-TLX, between the two conditions, $t(74)$ = -1.81, p = .074, d = -0.21. Still on a descriptive level, the *Physiology-based Adaptation*-condition produced more workload (M = 71.79, SD = 14.94) as compared to the control-condition(M = 66.83, SD = 17.86). There were no significant differences between conditions on the subscales, all d < .20.

In sum, we found in Experiment II that user expectations could be manipulated by a description of an adaptive system (H1). Participants expected superior performance when being assigned to the *Physiology-based Adaptation*-condition as compared to the *No Adaptivity*-condition. As indicated by the number of errors when solving the word puzzles, user performance was not affected by our sham-treatment (H2); however, as in Experiment I, we found a substantial correlation between task performance and user expectations, which gives partial support to H2. Much like in Experiment I, we found no effect of assistance on workload during completion of our task (H3).

6 DISCUSSION

Our study investigated if and how evaluation of AI systems are affected by a placebo effect. A perceived improvement occurred after using the AI system. While we have not found any indications of a placebo effect on objective metrics of task performance, our study shows how system descriptions change performance expectancy and how these affect common usability metrics, e.g. performance ratings of the UTAUT, for different AI systems.

The sustained belief of improvement after interaction constitutes the placebo effect. Considering that describing system functionality is common in HCI studies of AI, it is possible that some studies using subjective evaluations have fallen short on controlling for placebo effects in their study design. Our results should thus motivate future researchers to consider placebo control to ensure that system evaluations are not biased by user expectations.

We attribute our findings to three critical factors. First, the manipulating system descriptions manifests subjectively in performance expectations of the *Physiology-based Adaptation* (H1). These expectations decrease slightly but persist after the use of the system (placebo effect; H2). Second, there was no direct effect of manipulating performance expectations on task performance across both studies (H2). Notably, this is in line with the literature on placebo and cognitive enhancement. For instance, a sham sound treatment [89] or an expectation manipulation of receiving Ritalin, a potent drug that increases focus [66], only increased subjective expectations but not the objective performance metrics themselves. This notion of placebo effects only manifesting subjectively is also supported by meta-analysis that compared subjective and objective placebo-responses across 130 medical studies [44]. However, we found a relationship between performance expectations and the percentage of correct answers: The more an individual participant believed in the superiority of the *Physiology-based Adaptation*, the larger the performance gain when solving word puzzles. Third, we found in Experiment I that participants do not validate the system's functionality. In Experiment I, both allegedly adaptive systems were indeed judged to be adaptive. Thus, the possibility to validate the adaptive algorithm does not alter the formation of performance expectations. This is in line with the literature on automation-bias; the tendency to disregard contradictory information when being supported by automated decision making systems [14]. We did not find any effects on workload (H3).

6.1 Theoretical Implications

6.1.1 Technology Acceptance. Our data can be integrated into the current evaluation models of technology acceptance [15, 102, 103] based on three points. First, we found that the mere presence of a system description, albeit with no functionality, influences system evaluation. Therefore, system usability and ease of use on a behavioral level are not the only critical factors to technological acceptance; subjective perception must also be considered. Second, in terms of performance expectation, beneficial subjective judgment persisted after system use in both of our experiments. One could explain this continuance of belief in terms of confirmation bias [76]), which describes the propensity of participants to gather evidence for confirming a hypothesis. Here, the confirmatory evidence of receiving an easy puzzle appearing after making an error, could have influenced the judgment more than non-confirmatory evidence, i.e., of receiving no adaptivity at all. Therefore, in line with theories on technology acceptance, we can conclude that expectations can persist even if participants are presented with counterfactual information. Our experimental design was not intended to test expectation confirmation effects, hence our system descriptions were not constructed to implement judgments regarding satisfaction. Instead, we focused on the performance of the interaction with adaptive AI systems. Still, we observed decreases in performance expectations after interaction with the alleged adaptive systems. However, this decrease was largely uniform across all adaptivities. Third, we observed performance gains only in participants who expected these gains. Technology acceptance models such as the UTAUT [104] emphasize the role of subjective evaluation, as participants' demographics, prior experience, and so on can mitigate both performance expectations and their influence over user behavior. In sum, our findings can emphasize the role of subjective expectations in theories of technological acceptance and call for close control for these when evaluating "superior" technology.

6.1.2 Placebo Theory and Classification. In placebo research, there are two concurrent mechanisms for explaining placebo effects. Expectancy-oriented accounts suggest that placebo effects appear due to an increased belief in the efficacy of a treatment. Conditioned response accounts, on the other hand, describe the placebo effect on the level of previously obtained stimulus-response (e.g., between taking a pill and healing) associations. Our AI-based

adaptive system was entirely novel to the subjects, thus we argue that no stimulus-response associations could form. Therefore, our results agree with the former expectancy-based account of placebo effects.

The placebo literature in Medicine and Psychology emphasizes either physical artifacts (e.g. pills) or psychological treatment [17, 95]. However, placebo effects impact upon the confidence that can be placed in allegedly functional power-ups in a range of fields, for example, influencing gaming immersion when employing a phantom AI [18, 20] or control modules [100], or when cooperating with an assistive AI. Therefore, our literature survey and the present experiments can add a novel category to the placebo literature; namely, placebos introduced by digital artifacts. Our study is the first to investigate such a placebo in interaction with an alleged adaptive AI. Future research may investigate how digital placebos differ from physical or purely psychological placebo treatments about subjective and objective evaluation metrics.

6.2 Controlling for Placebo Effects

Adding digital placebos to the placebo literature would entail quantifying and experimentally controlling for placebo effects in HCI studies, much like in psychological and medical studies.

6.2.1 Control in Pharmacological Trials is not Applicable to HCI. Indeed, the placebo effect can be so prominent in medicine that many pharmacological trials typically compare a novel medical treatment (e.g., a new medicine or vaccine) against a placebo-control group (see Figure 1b) that receives only the placebo. This is done to ensure that participants cannot tell if they are in the experimental or control group, hence controlling user expectations before the actual treatment. User studies in HCI are confronted with larger obstacles in terms of placebo-control compared to pharmacological trials [34]. Participants in user studies can be aware of the novelty of the user interface they are assessing by knowing in which user group they are, i.e., assuming that the novel interface, that is different from a known one, can be better. To investigate the expectancy with the participants' subjective perception and objective performance, we measure the expectancies before and after using each *Adaptivity*.

6.2.2 Polling Expectations Prior and Post-Interaction. In this case, participants are unlikely to be naive or hold neutral expectations when presented with a novel and allegedly powerful technology such as AI that suggests improvement in terms of human capabilities. In other words, it is incorrect to assume an equivalence between control and experimental groups prior to interaction with intelligent and non-intelligent systems. Unfortunately, this casts doubt on the validity and robustness of user studies in that area that involve the interaction with intelligent technology. As far as we know, it is atypical to poll user expectations before and after a user study. Nonetheless, this should be a necessary consideration if a strong belief in the superiority of an intelligent system can generate favorable evaluations. Thus, in line with a recent discourse on research practices [9], HCI researchers must entertain the idea that many reported findings might be biased in favor of novel technologies.

6.2.3 Adding Placebo-Control Conditions. What are the practical implications of placebo effects for conducting user research? User studies should focus on normalizing and controlling the bias of user expectations. By introducing an active control group, or even a placebo-control to the study design, one can investigate whether the novel system exceeds the usability above and beyond the expectations. However, employing a placebo-control can be inefficient, unethical, or not possible, for example, when a placebo can be easily identified or when working with vulnerable patient samples requiring active treatment. In such cases, one can statistically control for the effect of expectations as presented in this study. This requires the quantification of expectations before the test itself with subsequent normalization of the bias afterward in the analysis.

6.2.4 Objective Measures are Less Susceptible to Placebo Effects. One could also try to resolve the issue by turning to so-called objective measures [87], where the participant is blind to the measurement rationale employed (e.g. an implicit association test) or physiological measures where the response is hard to modulate deliberately

(e.g. electroencephalography [70]). We do not think that this is a promising approach for two reasons. First, we found effects not only on self-reports but also on a behavioral level, albeit no main effect but only a correlation. The participants' performance expectancy improved when they had the assistance of a superior technology (i.e. *Physiology-based Adaptation*). Second, physiological responses can also be altered by the participants' expectations [11], hence explaining mixed results when researching implicit physiologically-adapting systems [96]. In line with placebo studies [50, 95], we recommend assessing the participant's expectations before use and compare them with the expectations after use. Correlating the UI superiority with the user's task performance can also serve as a pretest to anticipate the impact of expectations on the study results. One could also entertain the idea that qualitative methods may be superior when it comes to the bias of user expectations. A skilled interviewer may narrow down and identify the design aspects that might yield a favorable attitude towards the system or if a positive evaluation is only due to differential expectations. A qualitative interview before interaction with a UI could also normalize user expectations by explicitly addressing the role of novelty in evaluating a technological invention and thus making study participants aware of their potentially biased responses beforehand. Whether these strategies present a suitable solution to placebo-control needs to be investigated in future studies.

6.3 Exploiting the Placebo Effect

6.3.1 Persistence of Effects. Our results show a strong positive correlation between the description-induced subjective user expectations and the task performance during *Physiology-based Adaptation*. Here, we find that the narrative of the condition leads to a belief towards the utility of adaptivity, which reflects in improved performance compared to participants who do not believe in the superiority of the adaptivity. This shows that high user expectations indeed produce a placebo effect regarding the participant's perception (i.e., H1 and H2 are supported). This trend was clearly visible for the *Physiology-based Adaptation* condition, but not in the *Error-based Adaptation*. A potential reason for this is the participants' ability to self-validate their own provided input directly during the *Error-based Adaptation*. In contrast, *Physiology-based Adaptation* adds an encouraging non-transparent layer of improvement. Probing the users' beliefs after the experiment resulted in a lower score for the *Physiology-based Adaptation* compared to *Error-based Adaptation*, effectively reversing their belief in the superiority of the two adaptivities. This was revealed on the UTAUT questionnaire, that showed a significant difference between *Error-based Adaptation* and *Physiology-based Adaptation* on the "Performance Expectancy" scale. This finding validates our assumption that high user expectations before interaction with an implicitly adapting UI (i.e. *Physiology-based Adaptation*) lower one's belief in superiority more after interaction, compared to an explicit UI adaptation (i.e., *Error-based Adaptation*).

6.3.2 Designing for Placebo. We suggest, in line with [18, 33, 100], that shaping the user's expectations before interaction with a novel technology can create better user experiences. Thus, communication in the system's abilities and functioning, even if highly inaccurate, can potentially increase usability. This hypothesis on communication in technological innovation certainly deserves further investigation regarding the constraints and applicability of placebo effects in a user evaluation. Especially when considering that the *Physiology-based Adaptation* condition was judged to be less adaptive after interaction as compared to the *Error-based Adaptation*, this suggests that dis-confirmed expectations can potentially lead to unfavorable evaluations. Still, if carefully framed, a narrative may gauge the perceived functioning and utility of a system.

6.3.3 FATE: Fairness, Accountability, Transparency, and Explainability. We have not polled the perceived system complexity but assume that there are differences between the *Error-based Adaptation* and the *Physiology-based Adaptation* that could relate to differences in perceived transparency. While the *Error-based Adaptation* could be considered a rather trivial system of an error resulting in an easier item, the *Physiology-based Adaptation* could be considered a rather complex and less transparent system. For expectations to manifest in performance

ratings post-interaction a semi-accurate mental model needs to be established, as only then can performance expectations be confirmed or disconfirmed [15, 33, 98, 103]. Revisiting our results from this perspective can explain why the *Physiology-based Adaptation* system in Experiment I could not produce an expectancy effect or a placebo effect. Participants were unable to probe the system before doing the task, thus increasing the perceived transparency of the AI. In contrast, Experiment II was designed as a within-subjects study with a standardized system exploration, doubled number of trials and familiarity with the task due to repeated measurements. Providing participants with more information to build their individual mental model of the relatively complex *Physiology-based Adaptation* system could have increased perceived transparency. Making the inner working of the AI accessible is a core feature in strengthening the user's belief in performance gains which should be tested in future studies. Considering our results on the placebo effect for the *Error-based Adaptation* [33], post-hoc explanations of system decisions could increase performance ratings.

Our studies can also add to the literature on trust in automated systems. Overtrust refers to the notion that subjective system capabilities exceed objective system capabilities [58]. While these studies typically investigate trust with regard to functional systems, our studies show that non-functional systems can also provide a belief in performance capabilities. One could argue that this is negligible in practice or may even be desirable for technology adoption in private domains including video games [20], but could be detrimental for safety-critical contexts such as a semi-autonomous vehicle. Indeed, the literature on automation bias can show that contradictory information is not integrated in subsequent interaction when using automated systems [14, 92]. Hence, automation bias paired with increased expectations could foster an even stronger placebo effect of AI but also shift accountability from the user towards the autonomous system. Aligned with this, we know that user perceived fairness of AI is subjective [2]. Manipulating perceived fairness by changing system descriptions could likewise be possible.

6.4 Limitations and Future Work

6.4.1 Assessing Qualitative Data. Although we demonstrate the placebo effect induced by expectancy in two experiments, we have only found out a little about the factors that make up the placebo effect. Moving from a large-scale online study to a more controlled in-person study could allow for a closer examination of what factors contributed to participants' ratings of superior performance after interaction. We thus encourage future work that investigates the placebo effect in human-machine interaction to use mixed methods, thus supplementing quantitative data (e.g. physiological and behavioral data) with qualitative statements that can closely examine what made participants believe in the alleged system's functionality.

6.4.2 Providing Participants with Incentives. About 30% of the participants in each study had to be excluded from the analysis due to careless responding or lack of engagement in the study. While participants were paid for their participation, thoughtful responding was not incentivized. Garcia et al. [33] have used a consensus-oriented financial incentive, that is, they gave participants a bonus if they chose the system that was preferred by the majority. Future studies that evaluate placebo effects in HCI should follow this approach to use such a consensus-oriented financial incentive in system performance ratings which could promote careful evaluation and building of a representative mental model. Whether this weakens or strengthens the placebo effects needs to be investigated.

6.4.3 Long-Term Effects. Some limitations have to be taken into account. Our current results showed that our participants' positive expectations of the adaptivity systems decreased after experiencing them. However, we could not establish if they might be eradicated entirely or even turn negative. It would be necessary to perform studies on prolonged use to understand this better. A more thorough understanding of the placebo effect, the factors that mitigate it with system familiarity, and the time-scale of these interactions would help us better understand higher-level issues such as early technological adoption, acceptance, and disillusionment.

6.4.4 Nocebo Effect. In this study, the placebo was an alleged adaptive AI that should increase performance by selecting appropriate word puzzles and presenting them to the participant. Other studies in the domain of cognitive enhancement also use a so-called nocebo-condition, which is a sham treatment that shapes an expectation of impairing or even worsening task performance. In our case, this would add another condition to the experimental design in which an alleged challenging AI is used, i.e., one that suggests the most challenging word puzzles. This would allow researchers to closely see whether expectations of improvement and impairment similarly produce subjective performance expectations as well as gains or decreases in task performance. We are aware of one study that investigated biasing effects of AI descriptions. Describing art as being generated by AI prompted less favorable evaluations compared to descriptions of human-made artwork [85]. Thus, adding a nocebo-condition may allow for closer inspection of expectation effects in HCI-studies.

6.4.5 Prior User Experience. It is possible that our instructions on the alleged adaptivities could have directly contributed to the strength of the placebo effect. However, it is worth noting that this is an arbitrary decision that all HCI researchers have to make in introducing their systems to users. The current study is inconclusive for the specific factors that could have resulted in a larger placebo effect in the *Physiology-based Adaptation* condition. We believe that this is because participants had no access to verifying if the alleged system worked. An alternative explanation that is in line with placebo-conditioning accounts [95], could be that the *Physiology-based Adaptation* was less familiar to our participants than *Error-based Adaptation* systems. As far as the subjective results indicate, our participants considered both alleged systems to be equally novel.

6.4.6 Variation in Novelty. A valid criticism of our system descriptions could be that the *Error-based Adaptation* and the *Physiology-based Adaptation* condition differed with respect to novelty. This could have produced the difference between the conditions for the relation of task performance and expectations. We do not think this is a likely scenario as perceived novelty did not differ between the adaptive conditions. Both adaptive systems were considered to be equally novel. We did not control for individual differences (e.g. verbal fluency, stress) and how this could have impacted upon task performance across the test conditions. However, this study benefits from a large sample size, and participants were randomly assigned to the different adaptivities. Follow-up research could model individual variance in perceived novelty explicitly by performing a pre-study test for individual traits that could modulate the placebo effect and include it in regression analyses.

6.4.7 Generalizability to other HCI Studies. Finally, our study presents a simple interaction with technology that constrains the placebo effect to an *implicit* condition and to a set of participants who believed in the superiority of the assistant. Note that these characteristics are also relevant in medical and psychological placebo studies [3, 27]. Whether other HCI studies inherent characteristics such as heightened user expectations and *implicit* technology, and how placebo effects can unfold in response to other study characteristics still needs to be investigated. Nevertheless, our study could strengthen the argument that employing a placebo-control group is necessary to properly infer the superiority of a novel UI [96]. Our current recommendations are limited to the placebo effect on task performance and user acceptance. However, we plan to evaluate alternative metrics for their susceptibility as well. For this, it would be necessary to extensively evaluate the metrics that are currently used to determine technological adoption and deployment. This could lead to a more comprehensive framework for developing analytical tools that control user expectations in user studies.

7 CONCLUSION

Users approach technology with certain expectations, and researchers themselves will describe their technology prior to testing in ways that could bias expectations. This work shows that subjective expectations and task performance concerning alleged technological innovation can be affected by the system description. We report a placebo effect; a belief of system functionality induced by a sham-treatment after interaction, in HCI in the

case of collaboration with an AI. Our study suggests that placebo effects can be present in HCI. Thus, we believe that much like other fields of human-related research, such as Psychology, medicine, and sport sciences, HCI studies should consider placebo-control. We provide the source code of the application, the collected data, and the analysis scripts to foster further research in this area[13].

APPENDIX

We present the used questionnaires in the following. When referring to Likert items, we use a seven-point Likert scale[14].

A   VALIDATION QUESTIONS

We asked participants three validation questions in experiment I and experiment II after presenting the instructions and before the participants started to solve the word puzzles. We asked the validation questions to investigate if participants understood the instructions correctly. Participants who did not answer all three questions correctly were excluded from the analysis.

(1) What are the two conditions present in this study with regard to the adaptation of the task difficulty?
   (a) No support through adaptation and AI-based analysis of the camera image.
   (b) AI-based analysis of the camera image and adaptation by weather conditions.
   (c) Adaptation by analysis of surroundings and adaptation by the current location.
(2) What is analyzed by the AI during the study?
   (a) Daylight.
   (b) Pulse and emotional states through facial features using the camera image.
   (c) Surrounding objects in the image.
(3) What are the measured metrics used for?
   (a) Evaluation of the Internet connection.
   (b) Adaptation of the task difficulty.
   (c) Selection of correct browser settings.

B   CUSTOMIZED QUESTIONS

We asked the following custom questions using seven-point Likert items after each condition.

(1) The word puzzles adapted to my stress level.
(2) The word puzzles were easy to solve.
(3) I am using the same or a similar assistive system on a regular basis.

C   ADAPTED QUESTIONS

We adapted the Likert scales from Denisova et al. [18] to match the narrative of the task description. We used five out of six questions that were presented using seven-point Likert items. We omitted one question measuring the system intervention on a player's character.

---

[13]https://doi.org/10.17605/OSF.IO/W4Q6J - last access 2022-04-10
[14]1: Do not agree at all; 7: Absolutely agree

Table 4. Adapted UTAUT questions asked after every condition.

Performance Expectancy	
	I find the assistant useful in my daily life.
	Using the assistant increases my chances of achieving things that are important to me.
	Using the assistant helps me accomplish things more quickly.
	Using the assistant increases my productivity.
Effort Expectancy	
	Learning how to use the assistant was easy for me.
	My interaction with the assistant is clear and understandable.
	I find the assistant easy to use.
	It is easy for me to become skillful at using the assistant.
Hedonic Motivation	
	Using the assistant is fun.
	Using the assistant is enjoyable.
	Using the assistant is very entertaining.

(1) The word puzzles were generated according to my stress level.
(2) New word puzzles in the task appeared based on my stress level.
(3) The word puzzles were matched to my stress level.
(4) The behavior of the word puzzles changed when I was feeling bored or stressed.
(5) The word puzzles were not adapted sensibly to my stress level.

D   UTAUT

We used the three UTAUT items *Performance Expectancy*, *Effort Expectancy*, and *Hedonic Motivation* to measure the user expectancy and satisfaction [102]. We adapted the questions to meet the objectives of our study. The questions were asked after every condition (see Table 4).

E   WELLS ET AL.

We used all six question items used by Wells et al. [109]. We probed the participants after each condition using seven-point Likert items (see Table 5).

F   NASA-TLX

We used the NASA-TLX questionnaire to quantify the perceived task load for each adaptivity [38]. We calculated the raw NASA-TLX based on the six items ranging between one and twenty[15].

(1) Mental Demand: How mentally demanding was the task?
(2) Physical Demand: How physically demanding was the task?
(3) Temporal Demand: How hurried or rushed was the pace of the task?
(4) Performance: How successful were you in accomplishing what you were asked to do?[16]
(5) Effort: How hard did you have to work to accomplish your level of performance?
(6) Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

---

[15]1: Very Low; 20: Very High
[16]1: Perfect; 20: Failure

Table 5. Questions about the effect of perceived novelty on the adoption of information technology innovations by Wells et al.

Personal Innovativeness	
	If I heard about a new information technology, I would look for ways to experiment with it.
	Among my peers, I am usually the first to try out new information technologies.
	I like to experiment with new information technologies.
Novelty	
	I found using the assistant to be a novel experience.
	Using the assistant is new and refreshing.
	The assistant represents a neat and novel way of adapting the task difficulty.
Attitude	
	Overall, it is a good idea to use the assistant.
	Overall, it is wise to use the assistant.
	Overall, it is effective to use the assistant.
Overall Reward	
	On the whole, considering all sorts of factors combined, it is rewarding to sign up for and use the assistant.
	Using the assistant will be rewarding.
Behavioral Intention	
	I intend to use the assistant to adapt the task difficulty.
	My intention is to use the assistant instead of other non-adaptive systems.
	It is my intention to use the assistant at home or work.
Overall Risk	
	On the whole, considering all sorts of factors combined, it is risky to sign up for and use the assistant.
	Using the assistant would be risky.
	Using the assistant would add great uncertainty to my interaction.

REFERENCES

[1] Dennis Ang and Alex Mitchell. 2017. Comparing Effects of Dynamic Difficulty Adjustment Systems on Video Game Experience. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Amsterdam, The Netherlands) *(CHI PLAY '17)*. Association for Computing Machinery, New York, NY, USA, 317–327. https://doi.org/10.1145/3116595.3116623

[2] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35, 3 (2020), 611–623. https://doi.org/10.1007/s00146-019-00931-w

[3] Henry K. Beecher. 1955. The powerful placebo. *Journal of the American Medical Association* 159, 17 (12 1955), 1602–1606. https://doi.org/10.1001/jama.1955.02960340022006 arXiv:https://jamanetwork.com/journals/jama/articlepdf/303530/jama_159_17_006.pdf

[4] T. Bentley. 2000. Biasing web site user evaluations: A study. In *Proceedings of the Conference of the Computer Human Interaction Special Interest Group of the Ergonomics Society of Australia (OzCHI2000)*. 130–134.

[5] David Bethge, Thomas Kosch, Tobias Grosse-Puppendahl, Lewis L. Chuang, Mohamed Kari, Alexander Jagaciak, and Albrecht Schmidt. 2021. *VEmotion: Using Driving Context for Indirect Emotion Prediction in Real-Time.* Association for Computing Machinery, New York, NY, USA, 638–651. https://doi.org/10.1145/3472749.3474775

[6] Walter Boot, Daniel Simons, Cary Stothart, and Cassie Stutts Berry. 2013. The Pervasive Problem With Placebos in Psychology Why Active Control Groups Are Not Sufficient to Rule Out Placebo Effects. *Perspectives on Psychological Science* 8 (07 2013), 445–454. https://doi.org/10.1177/1745691613491271

[7] Lionel Briand, Davide Falessi, Shiva Nejati, Mehrdad Sabetzadeh, and Tao Yue. 2014. Traceability and SysML Design Slices to Support Safety Inspections: A Controlled Experiment. *ACM Trans. Softw. Eng. Methodol.* 23, 1, Article 9 (Feb. 2014), 43 pages. https://doi.org/10.1145/2559978

[8] Dermot Browne. 2016. *Adaptive user interfaces.* Elsevier.

[9] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. 1–15. https://doi.org/10.1145/3290605.3300733

[10] Tathagata Chakraborti and Subbarao Kambhampati. 2019. (When) Can AI Bots Lie?. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19).* Association for Computing Machinery, New York, NY, USA, 53–59. https://doi.org/10.1145/3306618.3314281

[11] Luca Chittaro and Riccardo Sioni. 2014. Affective computing vs. affective placebo: Study of a biofeedback-controlled game for relaxation training. *International Journal of Human-Computer Studies* 72, 8 (2014), 663 – 673. https://doi.org/10.1016/j.ijhcs.2014.01.007 Designing for emotional wellbeing.

[12] Michael Correll. 2020. What Do We Actually Learn from Evaluations in the "Heroic Era" of Visualization? arXiv:2008.11250 [cs.HC]

[13] Jean Costa, Alexander T. Adams, Malte F. Jung, François Guimbretière, and Tanzeem Choudhury. 2016. EmotionCheck: Leveraging Bodily Signals and False Feedback to Regulate Our Emotions. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16).* Association for Computing Machinery, New York, NY, USA, 758–769. https://doi.org/10.1145/2971648.2971752

[14] Mary L Cummings. 2017. Automation bias in intelligent time critical decision support systems. In *Decision Making in Aviation.* Routledge, 289–294.

[15] Fred D. Davis. 1985. *A technology acceptance model for empirically testing new end-user information systems: Theory and results.* Ph.D. Dissertation. Massachusetts Institute of Technology.

[16] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (1989), 319–340. http://www.jstor.org/stable/249008

[17] Peter J de Jong, Robert van Baast, Amoud Arntz, and Harald Merckelbach. 1996. The placebo effect in pain reduction: the influence of conditioning experiences and response expectancies. *International journal of behavioral medicine* 3, 1 (1996), 14–29. https://doi.org/10.1207/s15327558ijbm0301_2

[18] Alena Denisova and Paul Cairns. 2015. The placebo effect in digital games: Phantom perception of adaptive artificial intelligence. In *Proceedings of the 2015 annual symposium on computer-human interaction in play.* 23–33.

[19] Alena Denisova and Paul Cairns. 2019. Player experience and deceptive expectations of difficulty adaptation in digital games. *Entertainment Computing* 29 (2019), 56 – 68. https://doi.org/10.1016/j.entcom.2018.12.001

[20] Alena Denisova and Eliott Cook. 2019. Power-Ups in Digital Games: The Rewarding Effect of Phantom Game Elements on Player Experience. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (Barcelona, Spain) *(CHI PLAY '19).* Association for Computing Machinery, New York, NY, USA, 161–168. https://doi.org/10.1145/3311350.3347173

[21] Peter A. Dinda, Gokhan Memik, Robert P. Dick, Bin Lin, Arindam Mallik, Ashish Gupta, and Samuel Rossoff. 2007. The User in Experimental Computer Systems Research. In *Proceedings of the 2007 Workshop on Experimental Computer Science* (San Diego, California) *(ExpCS '07).* Association for Computing Machinery, New York, NY, USA, 10–es. https://doi.org/10.1145/1281700.1281710

[22] Luís Duarte and Luís Carriço. 2012. "Blue Pill or Red Pill?": Placebo Effect and the Outcome on Physiological Player Performance Metrics. In *Proceedings of the 4th International Conference on Fun and Games* (Toulouse, France) *(FnG '12).* Association for Computing Machinery, New York, NY, USA, 93–96. https://doi.org/10.1145/2367616.2367627

[23] Luis Duarte and Luis Carriço. 2013. The cake can be a lie: placebos as persuasive videogame elements. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems.* 1113–1118.

[24] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19).* Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312787

[25] Paul Ekman and Wallace V Friesen. 2003. *Unmasking the face: A guide to recognizing emotions from facial clues.*

[26] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21).* Association for Computing Machinery, New York, NY, USA, 754–768. https://doi.org/10.1145/3472749.3474784

[27] Paul Enck, Ulrike Bingel, Manfred Schedlowski, and Winfried Rief. 2013. The placebo response in medicine: minimize, maximize or personalize? *Nature reviews Drug discovery* 12, 3 (2013), 191–204.

[28] Ayman Farahat. 2013. How effective is targeted advertising?. In *2013 American Control Conference.* 6014–6021. https://doi.org/10.1109/ACC.2013.6580780

[29] Damien G Finniss, Ted J Kaptchuk, Franklin Miller, and Fabrizio Benedetti. 2010. Biological, clinical, and ethical advances of placebo effects. *The Lancet* 375, 9715 (2010), 686–695.

[30] Magne Arve Flaten and Terry D Blumenthal. 1999. Caffeine-associated stimuli elicit conditioned responses: an experimental model of the placebo effect. *Psychopharmacology* 145, 1 (1999), 105–112.

[31] Markus Funk, Tilman Dingler, Jennifer Cooper, and Albrecht Schmidt. 2015. Stop Helping Me - I'm Bored! Why Assembly Assistance Needs to Be Adaptive. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (Osaka, Japan) *(UbiComp/ISWC'15 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1269–1273. https://doi.org/10.1145/2800835.2807942

[32] Krzysztof Z. Gajos, Mary Czerwinski, Desney S. Tan, and Daniel S. Weld. 2006. Exploring the Design Space for Adaptive Graphical User Interfaces. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Venezia, Italy) *(AVI '06)*. Association for Computing Machinery, New York, NY, USA, 201–208. https://doi.org/10.1145/1133265.1133306

[33] Pedro García García, Enrico Costanza, Jhim Verame, Diana Nowacka, and Sarvapali D. Ramchurn. 2021. Seeing (Movement) is Believing: The Effect of Motion on Perception of Automatic Systems Performance. *Human–Computer Interaction* 36, 1 (2021), 1–51. https://doi.org/10.1080/07370024.2018.1453815 arXiv:https://doi.org/10.1080/07370024.2018.1453815

[34] Andrew L Geers, Paul E Weiland, Kristin Kosbab, Sarah J Landry, and Suzanne G Helfer. 2005. Goal activation, expectations, and the placebo effect. *Journal of personality and social psychology* 89, 2 (2005), 143.

[35] Asma Ghandeharioun and Rosalind Picard. 2017. *BrightBeat: Effortlessly Influencing Breathing for Cultivating Calmness and Focus.* Association for Computing Machinery, New York, NY, USA, 1624–1631. https://doi.org/10.1145/3027063.3053164

[36] K. J. Gilhooly and C. E. Johnson. 1978. Effects of Solution Word Attributes on Anagram Difficulty: A Regression Analysis. *Quarterly Journal of Experimental Psychology* 30, 1 (1978), 57–70. https://doi.org/10.1080/14640747808400654

[37] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909

[38] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139 – 183. https://doi.org/10.1016/S0166-4115(08)62386-9

[39] Jan Hartmann, Antonella De Angeli, and Alistair Sutcliffe. 2008. Framing the user experience: information biases on website quality judgement. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 855–864.

[40] Matthew J. Helm, Andrew M. Swiergosz, Heather S. Haeberle, Jaret M. Karnuta, Jonathan L. Schaffer, Viktor E. Krebs, Andrew I. Spitzer, and Prem N. Ramkumar. 2020. Machine learning and artificial intelligence: Definitions, applications, and future directions. *Current reviews in musculoskeletal medicine* 13, 1 (2020), 69–76.

[41] Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On Being Told How We Feel: How Algorithmic Sensor Feedback Influences Emotion Perception. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 114 (Sept. 2018), 31 pages. https://doi.org/10.1145/3264924

[42] Kasper Hornbæk and Effie Lai-Chong Law. 2007. Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 617–626.

[43] Kasper Hornbæk. 2006. Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies* 64, 2 (2006), 79 – 102. https://doi.org/10.1016/j.ijhcs.2005.06.002

[44] Asbjørn Hróbjartsson and Peter C. Gøtzsche. 2001. Is the Placebo Powerless? *New England Journal of Medicine* 344, 21 (2001), 1594–1602. https://doi.org/10.1056/NEJM200105243442106 arXiv:https://doi.org/10.1056/NEJM200105243442106 PMID: 11372012.

[45] Eva Hudlicka and Michael D Mcneese. 2002. Assessment of user affective and belief states for interface adaptation: Application to an Air Force pilot task. *User Modeling and User-Adapted Interaction* 12, 1 (2002), 1–47.

[46] Aviarini Indrati, Edi Minaji, Sugiharti Binastuti, and Philipus Dwi Raharjo. 2014. Comparison of Model Unified Theory of Acceptance and Use Technology (UTAUT) And Technology Acceptance Model (TAM) for Internet Adoption of Credit Union Staff. In *The First International Credit Union Conference on Social Micro*.

[47] Myounghoon Jeon. 2012. A Systematic Approach to Using Music for Mitigating Affective Effects on Driving Performance and Safety. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (Pittsburgh, Pennsylvania) *(UbiComp '12)*. Association for Computing Machinery, New York, NY, USA, 1127–1132. https://doi.org/10.1145/2370216.2370455

[48] Ted J Kaptchuk. 1998. Powerful placebo: the dark side of the randomised controlled trial. *The lancet* 351, 9117 (1998), 1722–1725. https://doi.org/10.1016/S0140-6736(97)10111-8

[49] Matthew Kay and Jacob O Wobbrock. 2016. Package 'ARTool'. *CRAN Repository* (2016), 1–13.

[50] Irving Kirsch. 1999. Specifying Nonspecifics: Psychological. *The placebo effect: An interdisciplinary exploration* 8 (1999), 166.

[51] Thomas Kosch, Markus Funk, Albrecht Schmidt, and Lewis L. Chuang. 2018. Identifying Cognitive Assistance with Mobile Electroencephalography: A Case Study with In-Situ Projections for Manual Assembly. *Proc. ACM Hum.-Comput. Interact.* 2, EICS, Article 11 (jun 2018), 20 pages. https://doi.org/10.1145/3229093

[52] Thomas Kosch, Mariam Hassib, Robin Reutter, and Florian Alt. 2020. Emotions on the Go: Mobile Emotion Assessment in Real-Time Using Facial Expressions. In *Proceedings of the International Conference on Advanced Visual Interfaces* (Salerno, Italy) *(AVI '20)*. Association for Computing Machinery, New York, NY, USA, Article 18, 9 pages. https://doi.org/10.1145/3399715.3399928

[53] Thomas Kosch, Mariam Hassib, Paweł W. Wozniak, Daniel Buschek, and Florian Alt. 2018. Your Eyes Tell: Leveraging Smooth Pursuit for Assessing Cognitive Workload. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174010

[54] Thomas Kosch, Jakob Karolus, Havy Ha, and Albrecht Schmidt. 2019. Your Skin Resists: Exploring Electrodermal Activity As Workload Indicator During Manual Assembly. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (Valencia, Spain) *(EICS '19)*. ACM, New York, NY, USA, Article 8, 5 pages. https://doi.org/10.1145/3319499.3328230

[55] Thomas Kosch, Albrecht Schmidt, Simon Thanheiser, and Lewis L. Chuang. 2020. One Does Not Simply RSVP: Mental Workload to Select Speed Reading Parameters Using Electroencephalography. 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/10.1145/3313831.3376766

[56] Nataliya Kosmyna, Franck Tarpin-Bernard, and Bertrand Rivet. 2015. Conceptual Priming for In-Game BCI Training. *ACM Trans. Comput.-Hum. Interact.* 22, 5, Article 26 (oct 2015), 25 pages. https://doi.org/10.1145/2808228

[57] John K. Kruschke. 2013. Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General* 142, 2 (2013), 573.

[58] Thomas Kundinger, Philipp Wintersberger, and Andreas Riener. 2019. (Over) Trust in Automated Driving: The Sleeping Pill of Tomorrow?. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[59] Dong-Heon Kwak, Philipp Holtkamp, and Sung S Kim. 2019. Measuring and controlling social desirability bias: Applications in information systems research. *Journal of the Association for Information Systems* 20, 4 (2019), 5.

[60] Louis Lasagna, Frederick Mosteller, John M von Felsinger, and Henry K Beecher. 1954. A study of the placebo response. *The American journal of medicine* 16, 6 (1954), 770–779.

[61] Eunjung Lee, Gyu-Wan Kim, Byung-Soo Kim, and Mi-Ae Kang. 2014. A Design Platform for Emotion-Aware User Interfaces. In *Proceedings of the 2014 Workshop on Emotion Representation and Modelling in Human-Computer-Interaction-Systems* (Istanbul, Turkey) *(ERM4HCI '14)*. Association for Computing Machinery, New York, NY, USA, 19–24. https://doi.org/10.1145/2668056.2668058

[62] Magdalena Lewandowska, Jacek Rumiński, Tomasz Kocejko, and Jędrzej Nowak. 2011. Measuring pulse rate with a webcam — A non-contact method for evaluating cardiac activity. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*. 405–410.

[63] Yang Li, Ranjitha Kumar, Walter S. Lasecki, and Otmar Hilliges. 2020. Artificial Intelligence for HCI: A Modern Approach. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3375147

[64] Yuyu Lin, Jiahao Guo, Yang Chen, Cheng Yao, and Fangtian Ying. 2020. It Is Your Turn: Collaborative Ideation With a Co-Creative Robot through Sketch. 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, 1–14. https://doi.org/10.1145/3313831.3376258

[65] Yabing Liu, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. 2011. Analyzing Facebook privacy settings: User expectations vs. reality. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC* (11 2011). https://doi.org/10.1145/2068816.2068823

[66] Alison Looby and Mitch Earleywine. 2011. Expectation to receive methylphenidate enhances subjective arousal but not cognitive performance. *Experimental and clinical psychopharmacology* 19, 6 (2011), 433.

[67] Antonio Luque-Casado, José C. Perales, David Cárdenas, and Daniel Sanabria. 2016. Heart rate variability and cognitive processing: The autonomic response to task demands. *Biological Psychology* 113 (2016), 83 – 90. https://doi.org/10.1016/j.biopsycho.2015.11.013

[68] Karen A. Matthews, Karen L. Woodall, and Michael T. Allen. 1993. Cardiovascular reactivity to stress predicts future blood pressure status. *Hypertension* 22, 4 (1993), 479–485. https://doi.org/10.1161/01.HYP.22.4.479

[69] Jaroslav Michalco, Jakob Grue Simonsen, and Kasper Hornbaek. 2015. An Exploration of the Relation Between Expectations and User Experience. *International Journal of Human-Computer Interaction* 31, 9 (2015), 603–617. https://doi.org/10.1080/10447318.2015.1065696 arXiv:https://doi.org/10.1080/10447318.2015.1065696

[70] David Millett. 2001. Hans Berger: From psychic energy to the EEG. *Perspectives in biology and medicine* 44, 4 (2001), 522–542.

[71] Ben Miroglio, David Zeber, Jofish Kaye, and Rebecca Weiss. 2018. *The Effect of Ad Blocking on User Engagement with the Web*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 813–821. https://doi.org/10.1145/3178876.3186162

[72] Guy Montgomery and Irving Kirsch. 1996. Mechanisms of placebo pain reduction: an empirical investigation. *Psychological science* 7, 3 (1996), 174–176.

[73] Richard D Morey, Jeffrey N Rouder, Tahira Jamil, and Maintainer Richard D Morey. 2015. Package 'bayesfactor'. *URLh http://cran/r-projectorg/web/packages/BayesFactor/BayesFactor pdf i (accessed 1006 15)* (2015).

[74] Sneha Narayan, Nathan TeBlunthuis, Wm Salt Hale, Benjamin Mako Hill, and Aaron Shaw. 2019. All Talk: How Increasing Interpersonal Communication on Wikis May Not Enhance Productivity. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 101 (Nov. 2019), 19 pages. https://doi.org/10.1145/3359203

[75] Divya Natesan, Morgan Walker, and Shannon Clark. 2016. Cognitive bias in usability testing. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, Vol. 5. SAGE Publications Sage CA: Los Angeles, CA, 86–88.

[76] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[77] M Ojanen et al. 1994. Can the true effects of exercise on psychological variables be separated from placebo effects? *International Journal of Sport Psychology* 25, 1 (1994), 63–80.

[78] Richard L Oliver. 1977. Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of applied psychology* 62, 4 (1977), 480.

[79] Anne Ottenbreit-Leftwich, Krista Glazewski, Minji Jeon, Cindy Hmelo-Silver, Bradford Mott, Seung Lee, and James Lester. 2021. *How Do Elementary Students Conceptualize Artificial Intelligence?* Association for Computing Machinery, New York, NY, USA, 1261. https://doi.org/10.1145/3408877.3439642

[80] Steffi Paepcke and Leila Takayama. 2010. Judging a Bot by Its Cover: An Experiment on Expectation Setting for Personal Robots. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (Osaka, Japan) *(HRI '10)*. IEEE Press, 45–52.

[81] Karthikeyan Palanisamy, Murguappan Muthusamy, and Yaacob Sazali. 2013. Detection of Human Stress Using Short-Term ECG and HRV Signals. *Journal of Mechanics in Medicine and Biology* 13, 02 (2013), 1350038. https://doi.org/10.1142/S0219519413500383 arXiv:https://doi.org/10.1142/S0219519413500383

[82] Jan Papoušek and Radek Pelánek. 2017. Should We Give Learners Control Over Item Difficulty?. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava, Slovakia) *(UMAP '17)*. Association for Computing Machinery, New York, NY, USA, 299–303. https://doi.org/10.1145/3099023.3099080

[83] Chris Piech, Mehran Sahami, Jonathan Huang, and Leonidas Guibas. 2015. Autonomously Generating Hints by Inferring Problem Solving Policies. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale* (Vancouver, BC, Canada) *(L@S '15)*. Association for Computing Machinery, New York, NY, USA, 195–204. https://doi.org/10.1145/2724660.2724668

[84] Ruchika Shaurya Prakash, Mariam A Hussain, and Brittney Schirda. 2015. The role of emotion regulation and cognitive control in the association between mindfulness disposition and stress. *Psychology and aging* 30, 1 (2015), 160. https://doi.org/10.1037/a0038544

[85] Martin Ragot, Nicolas Martin, and Salomé Cojean. 2020. AI-Generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence?. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3334480.3382892

[86] David Reiley. 2011. "Which Half Is Wasted?": Controlled Experiments to Measure Online-Advertising Effectiveness. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Diego, California, USA) *(KDD '11)*. Association for Computing Machinery, New York, NY, USA, 777. https://doi.org/10.1145/2020408.2020535

[87] Saul Rosenzweig. 1933. The experimental situation as a psychological problem. *Psychological Review* 40, 4 (1933), 337.

[88] Isa Rutten and David Geerts. 2020. Better Because It's New: The Impact of Perceived Novelty on the Added Value of Mid-Air Haptic Feedback. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[89] Katharina A Schwarz and Christian Büchel. 2015. Cognition and the placebo effect–dissociating subjective perception and actual performance. *PloS one* 10, 7 (2015), e0130492.

[90] Anil Shankar, Sushil J. Louis, Sergiu Dascalu, Linda J. Hayes, and Ramona Houmanfar. 2007. User-Context for Adaptive User Interfaces. In *Proceedings of the 12th International Conference on Intelligent User Interfaces* (Honolulu, Hawaii, USA) *(IUI '07)*. Association for Computing Machinery, New York, NY, USA, 321–324. https://doi.org/10.1145/1216295.1216357

[91] Maureen J. Simmonds and Dimitrios Zikos. 2014. Computer Games to Decrease Pain and Improve Mood and Movement. In *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments* (Rhodes, Greece) *(PETRA '14)*. Association for Computing Machinery, New York, NY, USA, Article 56, 4 pages. https://doi.org/10.1145/2674396.2674435

[92] Linda J Skitka, Kathleen Mosier, and Mark D Burdick. 2000. Accountability and automation bias. *International Journal of Human-Computer Studies* 52, 4 (2000), 701–717.

[93] Katta Spiel, Sven Bertel, and Fares Kayali. 2017. "Not Another Z Piece!": Adaptive Difficulty in TETRIS. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5126–5131. https://doi.org/10.1145/3025453.3025721

[94] Aaron Springer and Steve Whittaker. 2019. Progressive Disclosure: Empirically Motivated Approaches to Designing Effective Transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 107–120. https://doi.org/10.1145/3301275.3302322

[95] Steve Stewart-Williams and John Podd. 2004. The placebo effect: dissolving the expectancy versus conditioning debate. *Psychological bulletin* 130, 2 (2004), 324.

[96] Megan Strait, Cody Canning, and Matthias Scheutz. 2014. Reliability of NIRS-Based BCIs: a placebo-controlled replication and reanalysis of Brainput. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. 619–630.

[97] Iftekhar M. Tanveer, Ru Zhao, Kezhen Chen, Zoe Tiet, and Mohammed Ehsan Hoque. 2016. AutoManner: An Automated Interface for Making Public Speakers Aware of Their Mannerisms. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (Sonoma, California, USA) *(IUI '16)*. Association for Computing Machinery, New York, NY, USA, 385–396. https://doi.org/10.1145/2856767.2856785

[98] James YL Thong, Se-Joon Hong, and Kar Yan Tam. 2006. The effects of post-adoption beliefs on the expectation-confirmation model for information technology continuance. *International Journal of Human-computer studies* 64, 9 (2006), 799–810.

[99] Shari Trewin, Diogo Marques, and Tiago Guerreiro. 2015. Usage of subjective scales in accessibility research. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. 59–67.

[100] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[101] Viswanath Venkatesh and Fred D. Davis. 2000. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science* 46, 2 (2000), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926 arXiv:https://doi.org/10.1287/mnsc.46.2.186.11926

[102] Viswanath Venkatesh, James YL Thong, and Xin Xu. 2012. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. *MIS quarterly* (2012), 157–178.

[103] Viswanath Venkatesh, James YL Thong, and Xin Xu. 2016. Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the association for Information Systems* 17, 5 (2016), 328–376.

[104] Viswanath Venkatesh, James Y. L. Thong, and Xin Xu. 2012. Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly* 36, 1 (2012), 157–178. http://www.jstor.org/stable/41410412

[105] Sahar Voghoei, Navid Hashemi Tonekaboni, Delaram Yazdansepas, Saber Soleymani, Abolfazl Farahani, and Hamid R. Arabnia. 2020. Personalized Feedback Emails: A Case Study on Online Introductory Computer Science Courses. In *Proceedings of the 2020 ACM Southeast Conference* (Tampa, FL, USA) *(ACM SE '20)*. Association for Computing Machinery, New York, NY, USA, 18–25. https://doi.org/10.1145/3374135.3385274

[106] Eric-Jan Wagenmakers, Jonathon Love, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Ravi Selker, Quentin F Gronau, Damian Dropmann, Bruno Boutin, et al. 2018. Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic bulletin & review* 25, 1 (2018), 58–76.

[107] Rina R. Wehbe, Edward Lank, and Lennart E. Nacke. 2017. Left Them 4 Dead: Perception of Humans versus Non-Player Character Teammates in Cooperative Gameplay. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. Association for Computing Machinery, New York, NY, USA, 403–415. https://doi.org/10.1145/3064663.3064712

[108] John D Wells, Damon E Campbell, Joseph S Valacich, and Mauricio Featherman. 2010. The effect of perceived novelty on the adoption of information technology innovations: a risk/reward perspective. *Decision Sciences* 41, 4 (2010), 813–843.

[109] John D. Wells, Damon E. Campbell, Joseph S. Valacich, and Mauricio Featherman. 2010. The Effect of Perceived Novelty on the Adoption of Information Technology Innovations: A Risk/Reward Perspective. *Decision Sciences* 41, 4 (2010), 813–843. https://doi.org/10.1111/j.1540-5915.2010.00292.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5915.2010.00292.x

[110] Gesa Wiegand, Malin Eiband, Maximilian Haubelt, and Heinrich Hussmann. 2020. "I'd like an Explanation for That!"Exploring Reactions to Unexpected Autonomous Driving. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (Oldenburg, Germany) *(MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, Article 36, 11 pages. https://doi.org/10.1145/3379503.3403554

[111] Danielle Wilde and Kristina Andersen. 2010. Part Science Part Magic: Analysing the OWL Outcomes. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* (Brisbane, Australia) *(OZCHI '10)*. Association for Computing Machinery, New York, NY, USA, 188–191. https://doi.org/10.1145/1952222.1952262