"AI enhances our performance, I have no doubt this one will do the same": The Placebo Effect Is Robust to Negative Descriptions of AI

AGNES M. KLOFT*, Aalto University, Finland
ROBIN WELSCH*, Aalto University, Finland
THOMAS KOSCH, HU Berlin, Germany
STEEVEN VILLA, LMU Munich, Germany

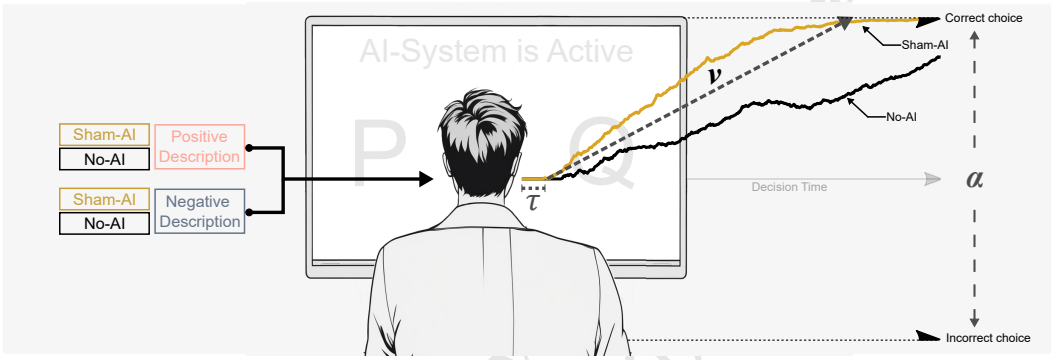

Fig. 1. Schematic representation of the drift-diffusion process of decision-making with an increased drift-rate v and a decreased non-decision time τ, for the sham-AI condition as compared to the no-AI condition. When using a sham-AI, participants accumulate information faster.

Heightened AI expectations facilitate performance in human-AI interactions through placebo effects. While lowering expectations to control for placebo effects is advisable, overly negative expectations could induce nocebo effects. In a letter discrimination task, we informed participants that an AI would either increase or decrease their performance by adapting the interface, but in reality, no AI was present in any condition. A Bayesian analysis showed that participants had high expectations and performed descriptively better irrespective of the AI description when a sham-AI was present. Using cognitive modeling, we could trace this advantage back to participants gathering more information. A replication study verified that negative AI descriptions do not alter expectations, suggesting that performance expectations with AI are biased and robust to negative verbal descriptions. We discuss the impact of user expectations on AI interactions and evaluation and provide a behavioral placebo marker for human-AI interaction.

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

Additional Key Words and Phrases: Placebo, Decision-making, Performance expectation

1 INTRODUCTION

Beliefs about Artificial intelligence (AI) fundamentally affect how we use this technology. The placebo effect of AI in Human-Computer Interaction (HCI) [40], inspired by medical research [6, 41, 41, 50, 70], documents that a sham-AI system can bring real subjective benefits accompanied by changes in behavior and physiology. Kosch et al. [40] argued that much like in the medical context, expectations about AI technology significantly influence study outcomes and thus undermine scientific evaluation if they are left uncontrolled.

*Shared first authorship: Both authors contributed equally to the paper

Prior research on placebo effects in HCI have been reported in gaming contexts, where fake power-up elements that make no difference to gameplay [19] and sham descriptions of AI adaptation increase game immersion [18]. In social media, sham control settings for a news feed can result in higher user satisfaction [74]. Kosch et al. [40] showed that a belief in receiving benefits from using an adaptive AI can improve performance. Villa et al. [78] could show that high expectations regarding sham-AI-based augmentation systems increase risky decision-making and affect information processing. Thus, AI technology can induce placebo effects that alter behavior through heightened positive expectations.

There are three major shortcomings in the placebo literature in HCI for AI. First, direct effects on a behavioral level are yet to be found [40, 78]. Second, it is unclear whether nocebo effects, low expectations impairing behavior, are equally influential as positive expectations based on verbal descriptions in HCI. Third, we lack a behavioral marker for effectively designing adaptive AI interfaces that enhance decision-making amidst placebo responses.

This paper provides an investigation of antecedents and consequences of AI's placebo effect in HCI. In an experimental study (N = 65), we examined the influence of negative and positive verbal AI descriptions. We analyzed the impact of expectations on decision-making in a letter discrimination task, with or without a sham-AI system.

First, in line with Kosch et al. [40], Villa et al. [78], we found a subjective placebo effect: participants retained belief in the sham-AI system's efficacy post-interaction. Second, we observed a main effect at the behavioral level. Utilizing a Bayesian cognitive model of decision-making revealed that participants gathered information faster and altered their response style, serving as a behavioral marker of the effect. Third, contrary to previous work [19, 40, 78], we found no effect of verbal descriptions. Participants were biased, expecting increased performance with AI, irrespective of the verbal descriptions (AI performance bias). We replicate this bias in an online study (N = 97).

Our results resonate with recent calls in HCI to control for placebo effects in evaluating AI systems [40, 78] and the power of AI narratives [14, 15, 38, 60]. We add to the literature an AI performance bias, which makes the AI's placebo effect robust to manipulations of verbal system descriptions. We provide a behavioral marker for the placebo response to AI, utilizing a cognitive model of decision-making to reduce bias in AI-driven user interfaces. We also discuss how, in a human-centered design process, the evaluation of AI must be done with user expectations in mind.

2 RELATED WORK

2.1 Expectations and the placebo effect of AI

People hold expectations with regard to AI. Survey findings show that fears about AI's disruptive impact outweigh excitement in the British public [14, 15]. This aligns with Sartori et al.'s report on the prevalence of 'AI anxiety' over perceived benefits [60]. Interestingly, this is not driven by actual technology but by narratives, as even science fiction portrayals contribute to the imbalance [31]. The narratives about AI can differ among stakeholders and change over time [8]. Indeed, national policies in countries like China, Germany, the USA, and France underscore AI's disruptive potential [4], and these narratives are widely impactful, affecting usage [8, 38]. Prior studies have explored key areas like transparency expectations [49, 52, 54], human-AI relationships [82], and autonomy [49], forming the basis for AI interface design. However, there's a gap in understanding expectations of human-AI interaction outcomes, such as task performance with AI support [40].

The placebo effect [5, 20, 33, 41, 53, 58] is not confined to medical contexts but also penetrates performance contexts like sports [7]. Here, an inert substance given to athletes can improve but also deteriorate performance [35]. While placebo effects of AI in HCI and their effect on performance have recently been studied [18, 40, 74, 78], there is very little knowledge on nocebo effects. In HCI,

a nocebo effect would negatively affect both performance and subjective metrics like usability or user experience [40]. For example, Ragot et al. [55] found that AI-generated art labeled as such was rated less favorably than if labeled as human-made. Halbhuber et al. [27] manipulated latency descriptions in gaming, showing that negative expectations reduced performance and user experience. **Thus, although first studies indicate the possibility of nocebo effects brought upon by technological artifacts, empirical studies directly leveling or even implementing negative expectations for AI are scarce.**

2.2 Decision-making with AI

Decision-making, a process shaped by expectations and perceptions, involves selecting from a range of options [67]. The Drift Diffusion Model (DDM) serves as a cognitive framework for understanding this process, describing it as evidence accumulation until a decision boundary (a correct vs. an incorrect answer) is reached [45, 51, 56, 57]. In its most basic form, the DDM models reaction times based on correct and incorrect responses in a random walk process toward a decision boundary, see Figure 1. For a binary decision task with equal probability, we can assume three parameters. A drift-rate v, indicating the speed of gathering information, a boundary separation α, reflecting a decision-strategy, and a non-decision time τ parameter, reflecting motor preparation and perceptual processes unrelated to decision-making[45]. The DDM's utility extends to HCI applications, including AI-enhanced interfaces, in various contexts like pedestrian crossing, social media, robot interactions or gaming[16, 17, 34, 43, 81].

Recent studies indicate that even sham adaptive AIs can influence user performance and risk-taking in decision-making [40, 78]. However, the cognitive mechanisms behind these effects remain unclear. Applying the DDM could potentially shed light on the cognitive basis of the placebo effect for adaptive AI systems. Considering previous studies Kosch et al. [40], Villa et al. [78], it appears plausible that the decision criterion may be affected by the implementation of positive expectations improving performance (more liberal decision-making with decreased α) and negative expectations (enlargement of α). **Consequently, users may make rapid, less accurate decisions when aided by an adaptive AI interface, or slower, more accurate decisions when the AI system potentially hampers their performance.**

3 RESEARCH MODEL AND HYPOTHESES

We conducted a mixed-design lab study with one between- and one within-subject factor, each with two levels. Two groups (between-subject) with different system descriptions, referred to as Description (the AI system is worsening performance and increases stress, referred to as negative verbal description condition vs. the AI system is enhancing performance and decreases stress, referred to as positive verbal description condition) were investigated. The within-subject factor for each group was the AI systems' Status (sham-AI condition vs. no-AI). The order of conditions was counterbalanced across participants.

We address the following research questions and hypotheses with this design:

4 METHOD

In the following, we motivate and document our methodological choices in realizing the study. The analysis with all associated measures can be found at https://osf.io/8q7t6/?view_only=3c67187606a84cd990aa4c6

4.1 Measures

4.1.1 Letter discrimination task. Two-alternative forced choice tasks, such as letter discrimination tasks, model simple decision-making and its underlying cognitive processes [46, 56, 72, 79]. In the task, participants must identify which of two letters, displayed on either side of a central target

Table 1. Research Questions and Hypotheses

Research Question	Hypotheses
RQ1: Do subjective ratings on performance and mental workload differ between negative and positive verbal descriptions (nocebo/placebo)?	**H1**: Lower subjective performance (**H1.1**) and higher mental workload (**H1.2**) in sham-AI with a negative description (nocebo) compared to no-AI. **H2**: Higher subjective performance (**H2.1**) and lower mental workload (**H2.2**) in sham-AI with a positive description (placebo) compared to no-AI.
RQ2: Do verbal descriptions of a sham-AI affect decision-making (e.g., in a letter discrimination task)?	**H3**: More conservative speed-accuracy trade-off in sham-AI with a negative description (nocebo) compared to no-AI. **H4**: More liberal speed-accuracy trade-off in sham-AI with a positive description (placebo) compared to no-AI.
RQ3: Do verbal descriptions of a sham-AI affect physiological indicators of cognition when compared to no-AI?	**H5**: Higher levels of physiological arousal (measured by mean tonic skin conductance level (SCL)) in sham-AI with a negative description (nocebo) compared to no-AI. **H6**: Lower levels of physiological arousal (measured by mean tonic SCL) in sham-AI with a positive description (placebo) compared to no-AI.

letter, matches the target. We used four letter pairs (E/F, P/R, C/G, Q/O), selected from Ratcliff and Rouder [56]. Each trial began with a fixation cross centrally displayed between the letters for a variable time (interstimulus interval, ISI), facilitating perception of the system's adaptability similar to [78]. Then, one of the letters was shown for 50.1 ms in the center of the screen [72]. After this, a randomly sketched line mask rotated by x * 360 degrees (x ∈ [0,1]) and mirrored (vertically and/or horizontally, or neither) was shown in place of the target letter for 1500 ms; see Figure 2. During the line mask presentation, the participants responded by pressing the left or the right arrow key (either index or middle finger). The first key press response during mask presentation time was recorded. The only critical change made to the task of Thapar et al. [72] was the randomly varying ISI. This was done to allow participants to track potential changes related to adaptation and should not affect task performance.

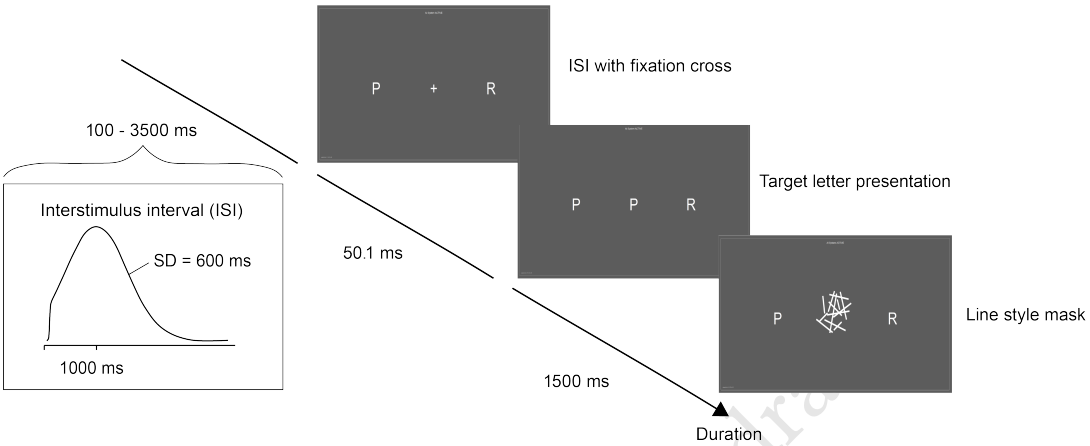

Fig. 2. Trial sequence during the letter discrimination task. The duration of the ISI followed a Gaussian distribution (M = 1000 ms, SD = 600 ms). Key responses (left or right arrow) were logged during the presentation of the mask.

Each participant underwent 400 trials derived from 2 Blocks × 100 trials of one random letter pair × 2 Status (sham-AI condition vs. no-AI). After each block, they were offered a short break.

4.1.2 *Questionnaires.*

Assessment of expectations. We measured user expectations of performance and how they persisted after the interaction. For overall performance expectations (judgments prior to interaction), we used four questions: A seven-point Likert item (1: Strongly disagree, and, 7 Strongly agree), *"I think I will perform better in the task with the AI system than in the task without the AI system."*, *"I will be [slower/faster] in the task with the AI system active than in the task with the AI system inactive."* on a slider (slower 0 to 100 faster) and two slider questions from zero to 200 asking participants the expected number of correct letter discriminations in each condition (*"Out of 200 actions, how many do you expect to get correct with/without the AI system active?"*). To evaluate judgments of performance following the interaction, identical questions phrased in the past tense were assessed. An additional questionnaire adapted from Villa et al. [78] was termed "System evaluation" and implemented to assess the participant's judgment of performance after the interaction, see Table 3.

Task load. To test H1 and H2 with respect to workload, we implemented the NASA-TLX [29], a well-established questionnaire [39], with six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. Participants rated each dimension on a scale of 1 to 20, with higher scores indicating higher task loads. We calculated the raw-score by summing up the item scores (Raw-TLX, [28]).

Additional Questionnaires. We assess user experience using the UEQ-S [62] (8 item pairs; Likert scale from -3 to +3) with its two dimensions, pragmatic quality and hedonic quality. For measuring Usability, we used an adapted version of the System Usability Scale (SUS) [10], changing "system" to "AI system", adding the synonym *awkward* for *cumbersome* [22], and computed the weighted sum score as an index of usability.

4.1.3 *EDA recording and pre-processing.*

Following the framework for EDA research in HCI [3], EDA was recorded using standard Ag/AgCl electrodes (24 mm surface diameter) placed on the

distal surfaces of the proximal phalanges of the index and middle fingers of the participant's non-dominant hand. Before testing, participants washed their hands with soap and cleaned the areas where the electrodes were placed using a 70% alcohol wipe. For data acquisition, we used the BITalino biomedical toolkit [26] to acquire the EDA signals via Bluetooth connection. The *OpenSignals (r)evolution* Python API Version 1.2.6[1] was set at sampling rate of 100 Hz. Time-series data were recorded using the Lab Streaming Layer (LSL)[2]. For offline data preprocessing, we used Neurokit toolbox [48]. After non-negative deconvolution analysis, we derived one metric of physiological arousal, the mean tonic SCL in each block.

4.2 Apparatus

The experiment was carried out using Chromium on a Linux (Ubuntu 22.04.2 LTS) laptop (Dell Latitude 7310) with an i5 (Intel Core i5-1031U) processor and 16GB of RAM. A separate monitor (HP E27uG4) displayed the paradigm with a screen size of 27 inches (2160px*1440px) and a refresh rate of 60 Hz. The monitor's position was adjusted according to the participant's eye level. Screen distance was roughly 60 cm (23,6 inch). We built a custom experiment that ran locally using JavaScript using the lab.js library version 20.2.4 [30].

4.3 Verbal Description

For the SHAM-AI condition, participants were informed that the AI system was continuously adapting the task difficulty based on their task performance and stress levels, monitored through electrodermal activity via electrodes. They were advised that recognizing these pace adjustments might take some time. In contrast, in the NO-AI condition, participants were informed that the AI system was not active and that the task pace was random.

Participants in the NEGATIVE VERBAL DESCRIPTION group were informed that the system had previously "decreased task performance" and resulted in an "elevation in stress" among users. Moreover, they were informed that the system was new and untested, thus making it **"unreliable"** and **"risky"** for use in real-world scenarios. In contrast, the participants in the POSITIVE VERBAL DESCRIPTION group were informed that the system had previously "enhanced task performance" while "reducing stress". They were also informed that the system was "cutting-edge", **"reliable"** and **"safe"** to use in real-world scenarios.

4.4 Participants

Participants were recruited through print advertisements in the [anonymized] area. Eligibility criteria included: no background in computer science, age above 18, self-reported normal or corrected-to-normal vision, no silver allergy, and no use of medication or history of epilepsy or other cognitive/motor impairments. The participants received 20 ANONYMIZED gift vouchers as compensation for their participation. The study was approved by an ethics committee (Grant Nr. <removed>).

We tested 66 participants in our study[3], excluding one for insufficient English proficiency. Our final sample size consisted of 65 participants (N = 65, ♂= 25, ♀= 40, zero non-binary or did not disclose) with an average age of 27.54 years (SD = 6.49) reporting an average technical competence of 4.82 (SD = 1.25) on a 1 (low) - 7 (high) Likert scale. To ensure that our samples are comparable, we checked their AI literacy using the Meta AI Literacy Scale[4] (MAILS) [12], the Checklist for Trust

[1] https://github.com/BITalinoWorld/revolution-python-api#bitalino-revolution-python-api/
[2] https://github.com/labstreaminglayer/
[3] Deviation from pre-registration see Table 7
[4] We only implemented items of factors loading onto the dimension "AI Literacy"

between People and Automation (TiA) [36] and the Subjective Information Processing Awareness Scale (SIPAS) [63–65]. We indeed found no differences as a function of Description (see Table 6).

4.5 Procedure

After consenting in line with the Declaration of Helsinki, the Bitalino device's electrodes were attached, and the device was activated and secured. The experimental program appeared on the screen. We then collected data on age, profession, handedness, caffeine or medication use, experimenter familiarity, and technical competence.

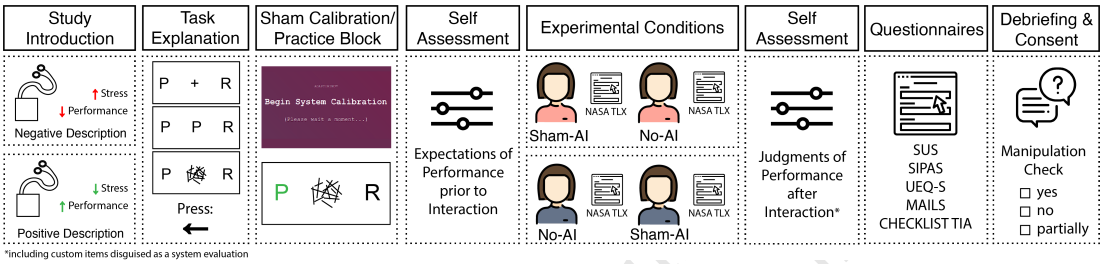

Fig. 3. Study Procedure: In this mixed-design study examining the induction of placebo and nocebo effects, participants were divided into two groups (Description), with each group receiving altered system descriptions. Participants in each group performed a letter discrimination task under two conditions (Status): in the sham-AI condition, they were informed the task pace was adjusted by an AI system based on their measured stress responses; in the No-AI condition, they were told adjustments in task pace were random. Before and after interacting with the sham-AI system, expectations on performance with and without the sham-AI system set as active were assessed. After the tasks and before debriefing, additional questionnaires assessing i.e., user experience and AI literacy were implemented. Ultimately, we revealed the manipulation and assessed the participants' belief in the manipulation.

Participants read an introductory text explaining the AI system and apparatus (see Figure 3). Depending on the Description assignment, the text included a positive or negative verbal description (Section 4.3) before interacting with the sham-AI. This was followed by a survey asking for information on the system being evaluated, see Villa et al. [78].

Before the task, participants completed 50 practice trials with visual feedback labeled as calibration. We then assessed their performance expectations with and without AI. Participants then performed the task in either the sham-AI or no-AI condition, based on group assignment. Task load was evaluated post-condition using TLX [29]. After both conditions, the AI system was further assessed (Section 4.1.2). Participants were then debriefed, re-consented, and their belief in the manipulation checked (Section 5.1) before thanking and compensating them.

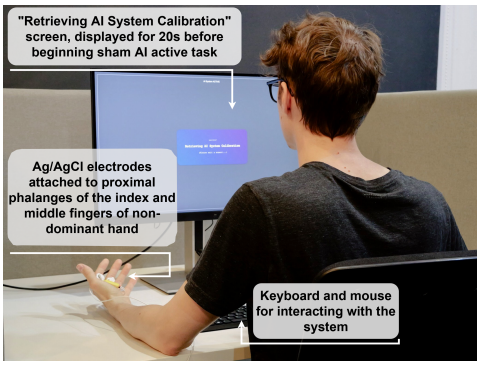

Fig. 4. The participants interacted with the system with their dominant hand using a keyboard and a mouse.

4.6 Bayesian Data Analysis and Inference

We adopted a Bayesian approach, utilizing Bayesian linear mixed models[5]. For parameter estimation, we used brms [11], a wrapper for the STAN-sampler [13] executed in R [71]. Two Hamilton-Monte-Carlo chains were computed, each with 4000-40,000 iterations and a 20% warm-up.

We then analyzed the posterior of the model. To investigate a parameter's distinguishability from zero, we utilized p_b, which resembles the classical p-value but quantifies the effect's likelihood of being zero or opposite [32, 68]. Effects with $p_b \leq 2.5\%$ were deemed distinguishable. We also calculated the 95% High-Density Interval (HDI) for each model parameter. For population-level effects in simple regression models, we set priors for regression parameters to one standard deviation of the outcome variable. All binary factors were effect coded (Time (pre/post): 1, -1; Status (sham-AI/no-AI): 1, -1; Description (negative/positive): 1, -1).

5 RESULTS

5.1 Manipulation Check

To the question *Did you believe that an AI system was implemented to adapt task pace?* with possible answers being *Yes, No* or *Partially*, 11 of 65 (16.92%; 6 of 33 negative description; 5 of 32 positive description) responded with "no" and did not believe in the system's capabilities. 26 out of 65 participants (40%; 13 for each description) participants reported some suspicion of the system's functionality. Thus, 28 of 65 participants fully believed in the implemented system.

5.2 Performance Expectations and Judgments of Performance

5.2.1 Subjective overall performance. To analyze expected overall performance, we centered the values by subtracting four points of the scale so that 0 indicates not favoring any condition and modeled overall performance estimates as a function of Time and Description[6]. Overall, participants were positive about the sham-AI, *Intercept* = 0.55 [0.29, 0.82], p_b = 0.00%. However, participants showed no difference in subjective performance before and after interaction with the sham-AI ($\tilde{b}_{\text{Time}}$ = 0.19 [-0.03, 0.41], p_b = 4.56%). There was no main effect of Description ($\tilde{b}_{\text{Description}}$= -0.23 [-0.50, 0.03], p_b = 4.3%) and no interaction effects ($\tilde{b}_{\text{Time} \times \text{Description}}$= 0.16 [-0.06, 0.38], p_b = 7.86%), see also Figure 5A.

[5]For a guide on Bayesian techniques, see[1, 11, 21, 25, 37, 61, 73, 76]
[6]Gaussian link-function with default priors.

5.2.2 Subjective estimated task speed. We computed a similar model to investigate participants' expected task speed. One participant was excluded due estimating the maximum of level 100 in all responses. Figure 5B shows the average expected speed across all conditions being positive, *Intercept* = 8.54 [5.42, 11.73], p_b = 0.00%). The participants believed to be faster with the sham-AI active before interacting with the system (M = 63.05, SD = 17.80) than after (M = 55.26, SD = 18.74). This difference could be distinguished from zero, $\tilde{b}_{\text{Time}}$ = 3.98 [0.99, 7.06] , p_b = 0.54% . We found no differences for Description $\tilde{b}_{\text{Description}}$ = -1.31 [-4.47, 1.88], p_b = 20.57% or interaction effects Description × Time $\tilde{b}_{\text{Description}}$ = -1.17 [-4.23, 1.88], p_b = 22.41% .

5.2.3 Subjective estimated number of correct responses. We expanded the statistical model to consider Status for estimated points in each condition; see also Figure 5C. Participants indicated that in the sham-AI condition (M = 143.16, SD = 30.42), they would score more points than in the no-AI condition (M = 130.31, SD = 32.35). This difference was not zero $\tilde{b}_{\text{Status}}$ = 6.41 [3.40, 9.40] , p_b = 0.00%. Participants believed to score more points before interacting with the sham-AI (M = 142.86, SD = 30.22) than after (M = 130.62, SD = 33.44), $\tilde{b}_{\text{Time}}$ = 6.16 [2.97, 9.30], p_b = 0.02%, resembling Kosch et al. [40]. We found no distinguishable effects for Description $\tilde{b}_{\text{Description}}$ = -2.05 [-9.32, 5.23], p_b = 29.09%, or any interaction effects p_b > 4.07%, see also Figure 5C.

Therefore, participants were biased toward a superior performance with AI even when given a negative verbal description of the system. We refer to this as AI performance bias.

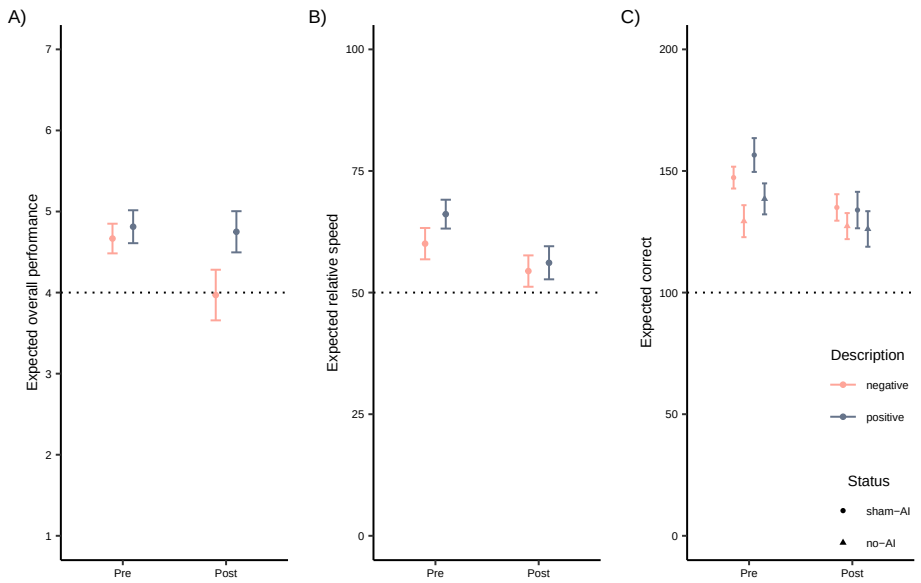

Fig. 5. A: Mean expected performance as a function of Time and Description. B: Mean expected relative speed as a function of Time and Description. C: Mean correct responses before and after interacting with the sham-AI system as a function of Time, Status and Description. Error bars denote +-1 standard error of the mean.

5.3 Performance data

We excluded 6 out of 65 participants (9.23%) from the behavioral data analysis as they did not comply with our task (percent correct <60% in one of the conditions or very large number of misses >35%). We deleted the first trial in each block along with too short responses by filtering reaction times (RT) under 150 ms (522 out of 23484; 2.22%)[7] and missed responses with RT > 1499 ms (32 out of 22930; 0.14%).

Table 2. Mean percent correct and reaction time (RT) for both correct and incorrect trials as a function of Status and Description

	Sham-AI			No-AI		
Description	Correct %	Correct RT	Incorrect RT	Correct %	Correct RT	Incorrect RT
Negative	91.11 (8.86)	586.80 (75.61)	719.16 (168.01)	90.35 (6.68)	596.56 (43.71)	739.32 (156.84)
Positive	89.12 (9.53)	596.77 (95.85)	718.99 (146.26)	88.15 (10.56)	598.49 (95.92)	716.86 (160.70)

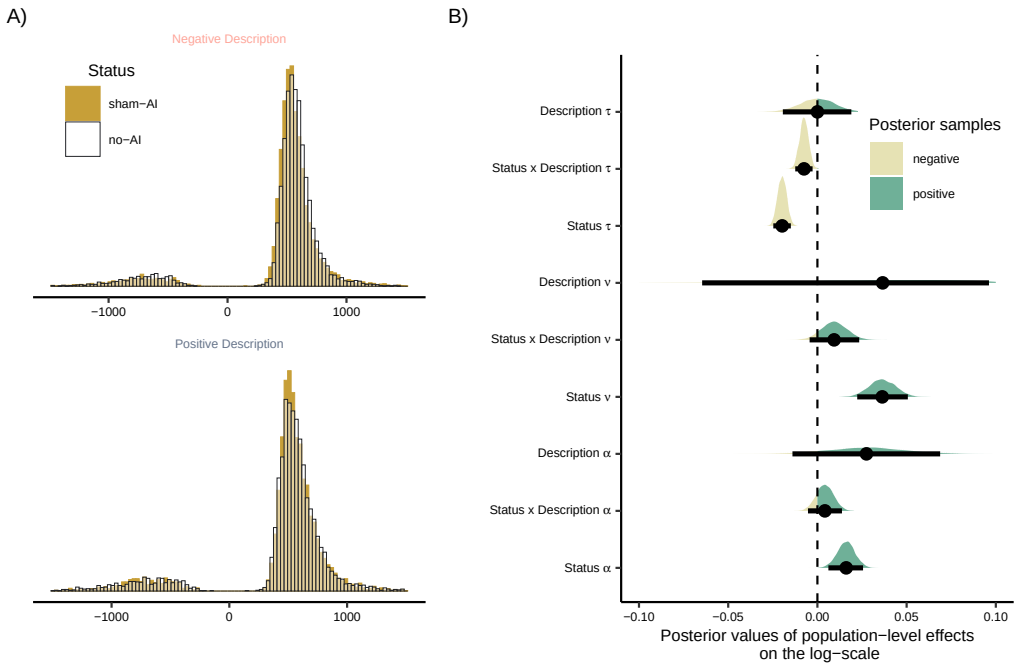

Fig. 6. A: Reaction time distribution as a function of Status (sham-AI vs. no-AI) and Description (incorrect trials are multiplied with -1). B: Posterior density plot for the parameter values for all population-level parameters 95% HDI. If the HDI does not cross the midline, p_b will be <2.5%.

Inspecting the RT of correct trials as a function of Status, we can see that RT distribution are slightly different, see Figure 6A, with sham-AI producing faster RTs for correct trials as compared to no-AI. This is also reflected in the mean of the distributions, see Table 2.

[7] Deviation from pre-registration see Table 7

We computed a DDM to test H3 & H4 on the reaction time data[8], see Figure 6A. A hierarchical form of this model was built accounting for inter-subject variability with a varying intercept and a population-level effect for each Status and an interaction term for Description for each τ, ν and α. We inspected the regression weights, see Figure 6B, for differences in Status and Description for boundary separation α, see Figure 7B, to see whether the difference in reaction time comes from a change in the participant's strategy, e.g., prioritizing speed over the accuracy. We found that in the sham-AI condition, participants had slightly larger boundary separation, α, making them slightly more conservative as compared to the no-AI condition. However, we also found that ν (drift rate) see Figure 7A, was higher for sham-AI as compared to no-AI. Thus, information accumulation was relatively faster in the sham-AI condition, see Figure 6B. Similarly, τ was also affected by Status with an interaction with Description qualifying the effect, see Table 8. Looking at Figure 6A and Figure 7C, we can see that the group with the negative description had a slightly earlier onset in RT. For all parameter values, see Table 8 or the mathematical formulation and priors, see Section B.

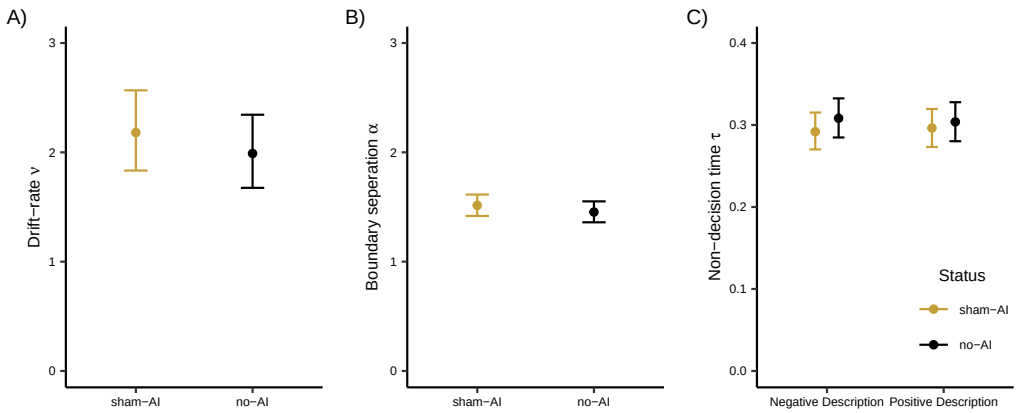

Fig. 7. A: Estimates with HDI 95% for the drift rate ν as a function of Status. B: Estimates with HDI 95% for boundary separation, α, as a function of Status. C: Estimates of non-decision time τ with HDI 95% as a function of Status and Description.

5.4 Workload and physiological arousal

Investigating H1.2, H2.2, H5 and H6, we computed a regression model for NASA-TLX raw data with Status and Description as predictors found[9]. We found no differences for Status $\tilde{b}_{\text{Status}}$ = -0.701 [-5.99, 4.26], p_b = 59.43%, Description $\tilde{b}_{\text{Description}}$ = -2.17 [-11.80, 7.88], p_b = 66.00% or interaction effects $\tilde{b}_{\text{Status} \times \text{Description}}$ = 1.55 [-6.08, 8.87], p_b = 65.55%. The EDA data largely resembled the TLX data. There was no effect of the Status, $\tilde{b}_{\text{Status}}$ = 0.05 [-0.24, 0.39], p_b = 38.62%, No effect of Description $\tilde{b}_{\text{Description}}$ = -0.68 [-1.57, 0.32], p_b = 6.88%, and no interaction effect, $\tilde{b}_{\text{Description} \times \text{Status}}$ = -0.30 [-0.94, 1.51], p_b = 32.5%.

[8]Deviation from pre-registration see Table 7
[9]Studentized link-function with priors scaled to one SD.

5.5 Usability and User Experience

Except for "The AI system made the task easier," which was viewed more favorably with a positive description, there were no significant differences in Description (Table 3). Participants slightly disagreed with "The task was easy" and were slightly negative about "The AI system improved my cognitive abilities." Yet, similar to Kosch et al. [40], Villa et al. [78], they agreed that the AI has future potential.

For UEQ-S scales, we found an overall positive user experience, with no group differences on hedonic or pragmatic attributes, with both having positive values indicating a positive user experience. SUS ratings indicated that the system was rated average in terms of usability unaffected by Description.

Table 3. Custom items for System evaluation were answered on a 7-point Likert scale (1 - strongly disagree; 7 - strongly agree). We estimate the difference towards a neutral value and compare the samples across Description. A neutral value for the custom items was 4, for the UEQ-S scales, was set to zero. We fitted a robust regression model for each comparison. For the SUS, the expected average is 68. Distinguishable effects from a neutral value (expected) or for each Description are marked with *. We used a studentized link-function with priors scaled to one *SD*

Item/scale	M_{neg} (*SD*)	M_{pos} (*SD*)	$\tilde{\Delta}_{\text{expected}}$[HDI 95%]	p_b	$\tilde{b}_{\text{Description}}$ [HDI 95%]	p_b
System evaluation						
The task was easy.	3.24 (1.62)	3.65 (1.80)	-0.56 [-0.99, -0.12]	0.58%*	-0.20 [-0.63, 0.23]	18.32%
The AI system						
- made the task easier.	3.36 (1.78)	4.28 (1.48)	-0.19 [-0.60, 0.22]	17.62%	-0.45 [-0.87, -0.05]	1.53%*
- made the task more enjoyable.	3.33 (1.83)	4.12 (1.49)	-0.28 [-0.70, 0.14]	9.70%	-0.39 [-0.80, 0.03]	3.51%
- made me more self-confident.	3.39 (1.49)	4.06 (1.77)	-0.28 [-0.70, 0.13]	9.05%	-0.32 [-0.73, 0.09]	6.48%
- made me more efficient.	3.87 (1.55)	4.25 (1.34)	0.06 [-0.30, 0.42]	36.26%	-0.16 [-0.52, 0.20]	18.61%
- improved my performance.	3.96 (1.55)	4.28 (1.39)	0.13 [-0.24, 0.50]	24.18%	-0.14 [-0.51, 0.22]	21.76%
- improved my cognitive abilities.	3.54 (1.39)	4.28 (1.34)	-0.37 [-0.71, -0.04]	1.55%*	-0.06 [-0.40, 0.28]	35.16%
- has a lot of potential for future development.	5.03 (1.15)	5.46 (1.13)	1.23 [0.95, 1.52]	0.00%*	-0.22 [-0.51, 0.07]	6.50%
UEQ-S-Pragmatic	0.53 (0.92)	0.92 (1.22)	0.75 [0.47, 1.01]	0.00%*	-0.20 [-0.46, 0.08]	7.44%
UEQ-S-Hedonic	0.73 (1.02)	0.80 (1.29)	0.77 [0.48, 1.06]	0.00%*	-0.04 [-0.32, 0.26]	39.88%
SUS	64.62 (13.86)	68.35 (17.50)	-1.11 [-5.05, 2.84]	28.87%	-2.04 [-5.93, 1.91]	15.3%

Note: In this figure, "neg" denotes a negative system description, while "pos" represents a positive one.

6 REPLICATION STUDY: POSITIVE EXPECTATIONS FOR NEGATIVE DESCRIPTIONS

To confirm the AI performance bias, we conducted an online replication study with negative system descriptions. We replicated the first part of the previous study[10]. As one could argue that participants did not comprehend the instructions, we set up the experiment to enforce comprehension of verbal descriptions. Participants read the negative system description and were asked to complete a comprehension check (Comprehension). The first group (N=51) completed the check and got no feedback on correctness, while the second group (N=46) had to answer all questions correctly (coded no: -1/yes: 1) to continue with the study. The system adaptation description was based on real-time video of their facial expressions, as per Kosch et al. [40] (no data was recorded). The data was z-standardized for modeling; regression weights, thus, show deviation from the mean, akin to Cohen's D. Finally, participants explained their point choices in an open text field. Participants were gathered through prolific.

[10]We used the same questions as for the assessment of expectations and judgments in the previous study

Table 4. Summary statistics for performance expectations as a function of Comprehension

Comprehension	Overall performance*	Expected task speed*	Δn Correct
No	5.11* (1.12)	69.48* (19.47)	-20.91* (30.60)
Yes	4.59* (0.96)	61.45* (17.70)	-2.06 (38.73)

*Note: Differences between groups are highlighted in the variable with a *, Means that are distinguishably more positive than their neutral value (4 for overall performance, 50 for task speed and r zero for nΔ for correct responses) are marked with *.*

6.1 Quantitative results

Table 4 shows all means of the subjective performance expectations for each group. Comprehension had an effect for overall performance, $\tilde{b}_{\text{Comprehension}}$= -0.26 [-0.47, -0.05] , p_b = 0.77% and expected task speed, $\tilde{b}_{\text{Comprehension}}$ = -4.02 [-7.73, -0.28], p_b = 1.71% but not for estimated correct, $\tilde{b}_{\text{Comprehension}}$ = 1.75 [-5.28, 8.89], p_b = 31.36%. Here, only a difference for Status emerged, $\tilde{b}_{\text{Status}}$ = 5.74 [2.16, 9.28] , p_b = 1.11% and an interaction effect $\tilde{b}_{\text{Status} \times \text{Comprehension}}$ = 5.74 [1.14, 8.28], p_b = 0.54%. Participants in the group without the enforced comprehension check estimated to answer more accurately with the sham-AI system active than without $p_b$diff =0.00%, while in the comprehension check group this difference was not present, $p_b$diff= 33.9%. Most importantly, participants were optimistic with regard to overall performance and expected speed, irrespective of Comprehension. Only for Δn for expected correct respinses, we find that the Comprehension leveled participants to neutral expectations.

6.2 Qualitative results

Participants were asked to explain the reasoning behind their responses after the estimation of performance. We analyzed all statements clustered for expected speed and overall performance. Two researchers independently performed a reflexive qualitative analysis, Braun et al. [9], of the statements, grouped them based on their semantic meaning, and finally agreed on a set of five categories, see Table 5.

Table 5. Subjective influence of the AI system on expected performance and speed before task completion: Number of statements per category

Category	Description	Statement Examples	Count (%) Performance	Speed
AI Trust	Trust and positive belief in capabilities of AI systems in general. Seeing AI as a powerful tool that ensures an advantage.	*AI models enhance our performances, so I have no doubt that this one will do the same.* (22P; P = 6) *Because I trust AI, IT IT FAST [sic] and quite reliable for most activities.* (11S; S = 68)	15 (15.96%)	12 (12.77%)
AI Assistance	AI is a helpful assistant that will facilitate task completion.	*I think the AI will assist me as it will be programmed to do the task, and I am not.* (35P; P = 6); *With the help of AI, I will be able to work fast because it will be assisting me rather than having to figure this out myself.* (37S; S = 85); *My effort and AI combined we will produce better results.* (40S; S = 99)	28 (29.79%)	40 (42.55%)
Uncertainty	Uncertainty toward the AI's systems influence on task completion.	*I don't know what to expect, really, maybe I could do better or not.* (18P; P = 4); *[...] I do not know the effects of the AI system on my performance yet.* (56S; S = 51)	31 (32.98%)	15 (15.96%)
Neutral	AI will neither have a positive or negative influence. AI won't make a difference in the task.	*I don't think there will be a large effect either way.* (2P; P = 4); *Because AI shouldn't have an effect on how I respond.* (15S; S = 56)	7 (7.45%)	9 (9.57%)
Self-Awareness	Self-reliance, and confidence in individual abilities, regardless of AI assistance, emphasizing autonomy and individual skill.	*Because I do not depend on enhancement to complete my tasks.* (7P; P = 6); *Cause I am a bit smarter for now than the AI system.* (39S; S = 99),	9 (9.57%)	8 (8.51%)
AI Antagonism	Lack of trust in the AI system, believing it will hamper performance, and skepticism towards AI's usefulness.	*As far as I understand, the AI will confuse me more than be of any help.* (60P; P = 4); *The AI might distract me and make me a little slower.* (9S; S = 27)	4 (4.26%)	10 (10.64%)

Note: The statements have been grammatically corrected to ensure good readability. Any quotes that remain unchanged are marked with [sic]. Each quote is followed by parentheses indicating the statement item number and whether the statement is related to expected performance (P) or speed (S). The number after the semicolon indicates the expected performance on a Likert Scale ranging from 1 (strongly disagree) to 7 (strongly agree), or expected speed, ranging from 1 (slower) to 100 (faster).

7 DISCUSSION

In this study, we set out to implement negative expectations and study the nocebo effect of AI (RQ1). However, we found that the placebo effect of AI in HCI [40] is robust to the manipulation of expectations by negative verbal description (contrary to H1.1 and H1.2). Even when we told participants that the AI system would make their task harder and more stressful, they still believed it would improve their performance. This was the same as for those who heard positive descriptions

of the AI (rendering H1, H3 & H5 void). We refer to this expectation of high performance as AI PERFORMANCE BIAS. We replicate this bias in a dedicated online study.

We found that heightened expectations (supporting H2.1.) carry over to the way participants make decisions (RQ2). Participants in the sham-AI condition responded slightly faster and more accurately when informed they were interacting with an adaptive AI system. Using the DDM model to analyze decision-making, we found that just believing an AI is involved can make participants gather information more quickly, respond more conservatively, and make them more alert (partial support for H4). We found no effects on workload or physiological arousal (no support for H2.1, H6).

7.1 AI performance bias as an antecedent of the placebo effect of AI

It appears that the prevailing positive perceptions about AI are influential enough to overshadow context-specific negative verbal descriptions. This could be due to participants bringing their daily experiences and narratives of AI into the evaluation, biasing both their subjective evaluations and behavior, see Table 5. From a mental model perspective [80], performance-reducing AI assistance may not fit into a coherent representation of human-AI interaction. It follows that the placebo mechanism for AI interfaces presented in the HCI literature is invalid [40, 78], as they focus on verbal system descriptions producing a placebo effect of AI. Based on our qualitative data, we follow that the effect is not specific to verbal descriptions of the system but may arise out of the socio-technical context as a function of the user's mental model.

The AI performance bias presents an intriguing contrast with Sartori and Bocca [60] findings on AI Anxiety. While individuals often express strong negative attitudes about AI replacing them in certain tasks, it appears that when humans and AI work together, even in a non-functional AI setting, joint performance is judged to be superior. Past studies have demonstrated that task performance can surpass individual AI or human performance in human-AI collaborations [23]. However, our findings shed new light on these findings. The human-AI performance gain may not arise from the summation of individual capabilities but also involves an elevation in human performance influenced by performance expectations. This suggests that (HCI) designers may harness the advantages of human-AI collaboration when focusing on systems that leverage a symbiotic relationship rather than fully automated tasks. However, future studies should explore not only the context of collaboration similar to Villa et al. [78] and Kosch et al. [40] but also consider human-AI competition.

7.2 The Impact of sham-AI on Decision-making

Villa et al. [78] explored the impact of the placebo effect on decision-making in risky situations. They found that individuals with high expectations of AI system support tended to take greater risks compared to those without AI assistance. This emphasizes how people's actions can be shaped by the narrative surrounding AI systems. In our study, we extended this research by investigating how positive and negative verbal descriptions affect decision-making processes. Our model showed that when people believed to have AI support, they gathered information faster than when not supported by AI. Yet, the type of narrative (positive or negative) did not have an impact on parameters in the DDM and thus, the underlying decision-making process. Prior research indicates that a participant's confidence can substantially influence the drift rate in a DDM [44, 47]. Therefore, it is possible that our findings can be explained by the participants feeling more confident when using the AI system. Also, we find a slightly more conservative decision boundary, with participants gathering more information until making a decision when supported with sham-AI. With AI support, participants might prioritize accuracy (a strategy that can be experimentally induced [69]), which also improves their overall performance. Lastly, sham-AI also shortened participants' non-decision time, indicating

they were in a more prepared state when making decisions, especially for negative descriptions. Note, however, that while some proponents associate a reduced non-decision time with better attention, as argued by Nunez et al. [51], or disinhibition [66], others have developed models without this parameter [77], as it is sensitive to contaminants. Thus, our computational model shows that the belief in using AI influenced participants' decision-making processes when interacting with a computing system.

7.3 Limitations & Implications

The study presents multiple limitations. First, by applying the social-affective perspective from Atlas [2] to our findings, it is evident that we did not account for the influence of emotions. While fostering a comfortable and friendly environment is commonly recommended in HCI evaluations [42, 59], prior research [24] has indicated that positive emotions can counteract the nocebo response in pain experiments. Positive affect could explain why we observed no nocebo effects. It is worth noting, however, that our study could not induce negative expectations to begin with, as confirmed by our validation study. Nonetheless, future research should take into account the impact of emotions during tests, perhaps by deliberately altering them, as suggested by Geers et al. [24].

In line with van Berkel and Hornbæk [75], we highlight two major domains of implications of our work. First, methodologically, given that a drift rate in the DDM can be estimated fast [79], the DDM could be used to compute a robust behavioral indicator of a placebo response for an AI interface. Second, it is crucial for the HCI community to understand that technology narratives can significantly bias AI performance expectations to the point where even negative descriptions cannot mitigate their influence on evaluation and interaction. For instance, positive expectations (placebo) may lead to overconfidence regarding the attributes of the system, such as its usability or user experience [40]. Our findings demonstrate that individuals tended to be overly confident about their performance. This could potentially mislead those evaluating the technology, fundamentally undermining the principles of human-centered design.

8 CONCLUSION

We found that even when we told participants to expect poor performance from a fake AI system, they still performed better and responded faster, showing a robust placebo effect. Contrary to previous work, this indicates that the placebo effect of AI is not easily negated by negative verbal descriptions, which raises questions about current methods for controlling for expectations in HCI studies. Additionally, the belief in having AI assistance facilitated decision-making processes, even when the narrative about AI was negative, thereby emphasizing that the influence of AI goes beyond simple narratives. This highlights the complexity and impact of AI narratives and suggests the need for a more nuanced approach in both research and practical user evaluation of AI.

REFERENCES

[1] Kevin Ackermans, Ellen Rusman, Rob Nadolski, Marcus Specht, and Saskia Brand-Gruwel. 2019. Video-or text-based rubrics: What is most effective for mental model growth of complex skills within formative assessment in secondary schools? *Computers in Human Behavior* 101 (Dec. 2019), 248–258. https://doi.org/10.1016/j.chb.2019.07.011

[2] Lauren Y Atlas. 2021. A social affective neuroscience lens on placebo analgesia. *Trends in Cognitive Sciences* 25, 11 (Nov. 2021), 992–1005. https://doi.org/10.1016/j.tics.2021.07.016

[3] Ebrahim Babaei, Benjamin Tag, Tilman Dingler, and Eduardo Velloso. 2021. A Critique of Electrodermal Activity Practices at CHI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(Chi '21)*. Association for Computing Machinery, New York, NY, USA, Article 177, 14 pages. https://doi.org/10.1145/3411764.3445370

[4] Jascha Bareis and Christian Katzenbach. 2022. Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science, Technology, & Human Values* 47, 5 (May 2022), 855–881. https://doi.org/10.1177/01622439211030007

[5] Ernest Edward Beckham. 1989. Improvement after evaluation in psychotherapy of depression: evidence of a placebo effect? *Journal of clinical psychology* 45, 6 (Nov. 1989), 945–950. https://doi.org/10.1002/1097-4679(198911)45:6<945::aid-jclp2270450620>3.0.co;2-2

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

[7] Christopher J Beedie, Damian A Coleman, and Abigail J Foad. 2007. Positive and negative placebo effects resulting from the deceptive administration of an ergogenic aid. *International journal of sport nutrition and exercise metabolism* 17, 3 (June 2007), 259–269. https://doi.org/10.1123/ijsnem.17.3.259

[8] Paolo Bory. 2019. Deep new: The shifting narratives of artificial intelligence from Deep Blue to AlphaGo. *Convergence* 25, 4 (Feb. 2019), 627–642. https://doi.org/10.1177/1354856519829679

[9] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. *Thematic Analysis BT - Handbook of Research Methods in Health Social Sciences*. Springer, Singapore, 843–860. https://doi.org/10.1007/978-981-10-5251-4_103

[10] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* 189 (Nov. 1995).

[11] Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software* 80, 1 (Aug. 2017), 1–28. https://doi.org/10.18637/jss.v080.i01

[12] Astrid Carolus, Martin Koch, Samantha Straka, Marc Latoschik, and Carolin Wienrich. 2023. MAILS – Meta AI Literacy Scale: Development and Testing of an AI Literacy Questionnaire Based on Well-Founded Competency Models and Psychological Change- and Meta-Competencies. (Feb. 2023). arXiv:2302.09319 [cs.AI]

[13] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A Probabilistic Programming Language. *Journal of Statistical Software* 76, 1 (Jan. 2017), 1–32. https://doi.org/10.18637/jss.v076.i01

[14] Stephen Cave, Kate Coughlan, and Kanta Dihal. 2019. "Scary Robots": Examining Public Responses to AI. 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, 331–337. https://doi.org/10.1145/3306618.3314232

[15] Stephen Cave and Kanta Dihal. 2019. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* 1 (Feb. 2019), 74–78. https://doi.org/10.1038/s42256-019-0020-9

[16] Francesco Chiossi, Luke Haliburton, Changkun Ou, Andreas Martin Butz, and Albrecht Schmidt. 2023. Short-Form Videos Degrade Our Capacity to Retain Intentions: Effect of Context Switching On Prospective Memory. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(Chi '23)*. Association for Computing Machinery, New York, NY, USA, Article 30, 15 pages. https://doi.org/10.1145/3544548.3580778

[17] Javier Corredor, Jorge Sofrony, and Angelika Peer. 2017. Decision-Making Model for Adaptive Impedance Control of Teleoperation Systems. *Institute of Electrical and Electronics Engineers (IEEE) Transactions on Haptics* 10, 1 (Jan. 2017), 5–16. https://doi.org/10.1109/toh.2016.2581807

[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* (London, United Kingdom) *(Chi Play '15)*. Association for Computing Machinery, New York, NY, USA, 23–33. https://doi.org/10.1145/2793107.2793109

[19] 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

[20] Nico J. Diederich and Christopher G. Goetz. 2008. The placebo treatments in neurosciences. *Neurology* 71, 9 (Aug. 2008), 677–684. https://doi.org/10.1212/01.wnl.0000324635.49971.3d

[21] Alan Dix. 2022. Bayesian statistics. In *Bayesian Methods for Interaction and Design*, John H. Williamson, Antti Oulasvirta, Per Ola Kristensson, and Nikola Banovic (Eds.). Cambridge University Press, 81–114. https://doi.org/10.1017/9781108874830.004

[22] Kraig Finstad. 2006. The System Usability Scale and Non-Native English Speakers. *Journal of User Experience* 1, 4 (Aug. 2006), 185–188.

[23] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. 2022. Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research* 33, 2 (June 2022), 678–696. https://doi.org/10.1287/isre.2021.1079 arXiv:https://doi.org/10.1287/isre.2021.1079

[24] Andrew L Geers, Shane Close, Fawn C Caplandies, Charles L Vogel, Ashley B Murray, Yopina Pertiwi, Ian M Handley, and Lene Vase. 2019. Testing a positive-affect induction to reduce verbally induced nocebo hyperalgesia in an experimental pain paradigm. *Pain* 160, 10 (Oct. 2019), 2290–2297. https://doi.org/10.1097/j.pain.0000000000001618

[25] Noa Gueron-Sela, Ido Shalev, Avigail Gordon-Hacker, Alisa Egotubov, and Rachel Barr. 2023. Screen media exposure and behavioral adjustment in early childhood during and after COVID-19 home lockdown periods. *Computers in*

Human Behavior 140 (March 2023), 11. https://doi.org/10.1016/j.chb.2022.107572

[26] José Guerreiro, Raúl Martins, Hugo Silva, André Lourenço, and Ana Fred. 2013. BITalino - A Multimodal Platform for Physiological Computing. In *Proceedings of the 10th International Conference on Informatics in Control, Automation and Robotics - Volume 1: ICINCO*. 500–506. https://doi.org/10.5220/0004594105000506

[27] David Halbhuber, Maximilian Schlenczek, Johanna Bogon, and Niels Henze. 2022. Better Be Quiet about It! The Effects of Phantom Latency on Experienced First-Person Shooter Players. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia* (Lisbon, Portugal) *(Mum '22)*. Association for Computing Machinery, New York, NY, USA, 172–181. https://doi.org/10.1145/3568444.3568448

[28] 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 (Oct. 2006), 904–908. https://doi.org/10.1177/154193120605000909

[29] 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

[30] Felix Henninger, Yury Shevchenko, Ulf Mertens, Pascal J. Kieslich, and Benjamin E. Hilbig. 2021. *lab.js: A free, open, online experiment builder*. https://doi.org/10.5281/zenodo.5233003

[31] Isabella Hermann. 2020. Beware of fictional AI narratives. *Nature Machine Intelligence* 2, 11 (Oct. 2020), 654–654. https://doi.org/10.1038/s42256-020-00256-0

[32] Herbert Hoijtink and Rens van de Schoot. 2018. Testing small variance priors using prior-posterior predictive p values. *Psychological Methods* 23, 3 (2018), 561–569. https://doi.org/10.1037/met0000131

[33] Asbjørn Hrõbjartsson and Peter Christian Gøtzsche. 2001. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *The New England journal of medicine* 344 21 (May 2001), 1594–602. https://doi.org/10.1056/nejm200105243442106

[34] Jie Huang, Wenhua Wu, Zhenyi Zhang, and Yutao Chen. 2020. A human decision-making behavior model for human-robot interaction in multi-robot systems. *Institute of Electrical and Electronics Engineers (IEEE) Access* 8 (Nov. 2020), 197853–197862. https://doi.org/10.1109/access.2020.3035348

[35] Philip Hurst, Lieke Schipof-Godart, Attila Szabo, John Raglin, Florentina Hettinga, Bart Roelands, Andrew Lane, Abby Foad, Damian Coleman, and Chris Beedie. 2020. The Placebo and Nocebo effect on sports performance: A systematic review. *European Journal of Sport Science* 20, 3 (Aug. 2020), 279–292. https://doi.org/10.1080/17461391.2019.1655098

[36] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (March 2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[37] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(Chi '16)*. Association for Computing Machinery, New York, NY, USA, 4521–4532. https://doi.org/10.1145/2858036.2858465

[38] William R. King and Jun He. 2006. A meta-analysis of the technology acceptance model. *Information & Management* 43, 6 (Sept. 2006), 740–755. https://doi.org/10.1016/j.im.2006.05.003

[39] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Comput. Surv.* 55, 13s, Article 283 (July 2023), 39 pages. https://doi.org/10.1145/3582272

[40] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2023. The Placebo Effect of Artificial Intelligence in Human–Computer Interaction. *ACM Trans. Comput.-Hum. Interact.* 29, 6, Article 56 (Jan. 2023), 32 pages. https://doi.org/10.1145/3529225

[41] 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 (June 1954), 770–779. https://doi.org/10.1016/0002-9343(54)90441-6

[42] Jonathan Lazar, Julio Abascal, Simone Barbosa, Jeremy Barksdale, Batya Friedman, Jens Grossklags, Jan Gulliksen, Jeff Johnson, Tom McEwan, Loïc Martínez-Normand, Wibke Michalk, Janice Tsai, Gerrit van der Veer, Hans von Axelson, Ake Walldius, Gill Whitney, Marco Winckler, Volker Wulf, Elizabeth F. Churchill, Lorrie Cranor, Janet Davis, Alan Hedge, Harry Hochheiser, Juan Pablo Hourcade, Clayton Lewis, Lisa Nathan, Fabio Paterno, Blake Reid, Whitney Quesenbery, Ted Selker, and Brian Wentz. 2016. Human–Computer Interaction and International Public Policymaking: A Framework for Understanding and Taking Future Actions. *Foundations and Trends in Human-Computer Interaction* 9, 2 (May 2016), 69–149. https://doi.org/10.1561/1100000062

[43] Byungjoo Lee, Sunjun Kim, Antti Oulasvirta, Jong-In Lee, and Eunji Park. 2018. Moving Target Selection: A Cue Integration Model. 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, 1–12. https://doi.org/10.1145/3173574.3173804

[44] Douglas G. Lee, Jean Daunizeau, and Giovanni Pezzulo. 2021. Evidence or Confidence: What Is Really Monitored during a Decision? *Psychonomic Bulletin & Review* (March 2021), 1–20. https://doi.org/10.3758/s13423-023-02255-9

[45] Veronika Lerche and Andreas Voss. 2018. Speed–accuracy manipulations and diffusion modeling: Lack of discriminant validity of the manipulation or of the parameter estimates? *Behavior Research Methods* 50 (March 2018), 2568–2585. https://doi.org/10.3758/s13428-018-1034-7

[46] Veronika Lerche, Andreas Voss, and Markus Nagler. 2017. How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods* 49 (April 2017), 513–537. https://doi.org/10.3758/s13428-016-0740-2

[47] Yaxin Liu and Stella F. Lourenco. 2022. Drift diffusion modeling informs how affective factors affect visuospatial decision making. *Journal of Vision* 22 (Dec. 2022), 3394. https://doi.org/10.1167/jov.22.14.3394

[48] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods* 53 (Feb. 2021), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

[49] Christian Meurisch, Cristina A. Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. 2020. Exploring User Expectations of Proactive AI Systems. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 146 (Dec. 2020), 22 pages. https://doi.org/10.1145/3432193

[50] Guy Montgomery and Irving Kirsch. 1996. Mechanisms of Placebo Pain Reduction: An Empirical Investigation. *Psychological Science* 7 (May 1996), 174–176. https://doi.org/10.1111/j.1467-9280.1996.tb00352.x

[51] Michael D. Nunez, Joachim Vandekerckhove, and Ramesh Srinivasan. 2017. How attention influences perceptual decision making: Single-trial EEG correlates of drift-diffusion model parameters. *Journal of Mathematical Psychology* 76 (Feb. 2017), 117–130. https://doi.org/10.1016/j.jmp.2016.03.003

[52] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2021. Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(Chi '21)*. Association for Computing Machinery, New York, NY, USA, Article 154, 15 pages. https://doi.org/10.1145/3411764.3445304

[53] Donald D. Price, Damien G. Finniss, and Fabrizio Benedetti. 2008. A Comprehensive Review of the Placebo Effect: Recent Advances and Current Thought. *Annual Review of Psychology* 59 (Jan. 2008), 565–590. https://doi.org/10.1146/annurev.psych.59.113006.095941

[54] Zoe A. Purcell, Mengchen Dong, Anne-Marie Nussberger, Nils Köbis, and Maurice Jakesch. 2023. Fears about AI-mediated communication are grounded in different expectations for one's own versus others' use. arXiv:2305.01670 [cs.HC]

[55] 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

[56] Roger Ratcliff and Jeffrey N. Rouder. 2000. A diffusion model account of masking in two-choice letter identification. *Journal of experimental psychology. Human perception and performance* 26, 1 (Feb. 2000), 127–40. https://doi.org/10.1037//0096-1523.26.1.127

[57] Roger Ratcliff and Philip L. Smith. 2010. Perceptual discrimination in static and dynamic noise: The temporal relation between perceptual encoding and decision making. *Journal of Experimental Psychology: General* 139, 1 (Feb. 2010), 70–94. https://doi.org/10.1037/a0018128

[58] K. Rickels, P. T. Hesbacher, C. C. Weise, B. Gray, and H. S. Feldman. 1970. Pills and improvement: A study of placebo response in psychoneurotic outpatients. *Psychopharmacologia* 16 (Jan. 1970), 318–328. https://doi.org/10.1007/bf00404738

[59] Jeffrey Rubin and Dana Chisnell. 2008. *Handbook of usability testing: How to plan, design, and conduct effective tests.* John Wiley & Sons.

[60] Laura Sartori and Giulia Bocca. 2023. Minding the gap(s): public perceptions of AI and socio-technical imaginaries. *Ai & Society* 38, 2 (April 2023), 443–458. https://doi.org/10.1007/s00146-022-01422-1

[61] Daniel J Schad, Michael Betancourt, and Shravan Vasishth. 2021. Toward a principled Bayesian workflow in cognitive science. *Psychological methods* 26, 1 (2021), 103.

[62] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2017. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence* 4 (2017), 103–108. https://doi.org/10.9781/ijimai.2017.09.001

[63] Tim Schrills and Thomas Franke. 2021. Subjective Information Processing Awareness Scale (SIPAS).

[64] Tim Schrills and Thomas Franke. 2023. How Do Users Experience Traceability of AI Systems? Examining Subjective Information Processing Awareness in Automated Insulin Delivery (AID) Systems. *ACM Trans. Interact. Intell. Syst.*

(March 2023). https://doi.org/10.1145/3588594
[65] Tim Schrills, Mourad Zoubir, Mona Bickel, Susanne Kargl, and Thomas Franke. 2021. Are Users in the Loop? Development of the Subjective Information Processing Awareness Scale to Assess XAI.
[66] Stefanie Schuch. 2016. Task inhibition and response inhibition in older vs. younger adults: A diffusion model analysis. *Frontiers in Psychology* 7 (2016), 1722.
[67] Eldar Shafir, Itamar Simonson, and Amos Tversky. 1993. Reason-based choice. *Cognition* 49, 1 (1993), 11–36. https://doi.org/10.1016/0010-0277(93)90034-S
[68] Haolun Shi and Guosheng Yin. 2020. Reconnecting p-value and Posterior Probability under One-and Two-sided Tests. *The American Statistician* (2020), 1–11.
[69] Jeffrey J Starns and Roger Ratcliff. 2010. The effects of aging on the speed–accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging* 25, 2 (2010), 377.
[70] Steve Stewart-Williams and John Podd. 2004. The placebo effect: dissolving the expectancy versus conditioning debate. *Psychological Bulletin* 130, 2 (2004), 324.
[71] R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[72] Anjali Thapar, Roger Ratcliff, and Gail McKoon. 2003. A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging* 18, 3 (2003), 415–429. https://doi.org/10.1037/0882-7974.18.3.415
[73] Rafal Urbaniak, Patrycja Tempska, Maria Dowgiałło, Michał Ptaszyński, Marcin Fortuna, Michał Marcińczuk, Jan Piesiewicz, Gniewosz Leliwa, Kamil Soliwoda, Ida Dziublewska, Nataliya Sulzhytskaya, Aleksandra Karnicka, Paweł Skrzek, Paula Karbowska, Maciej Brochocki, and Michał Wroczyński. 2022. Namespotting: Username toxicity and actual toxic behavior on Reddit. *Computers in Human Behavior* 136 (2022), 107371. https://doi.org/10.1016/j.chb.2022.107371
[74] 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* (Montreal QC, Canada) *(Chi '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173590
[75] Niels van Berkel and Kasper Hornbæk. 2023. Implications of Human-Computer Interaction Research. *Interactions* 30, 4 (June 2023), 50–55. https://doi.org/10.1145/3600103
[76] Rens van de Schoot, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G. Tadesse, Marina Vannucci, Andrew Gelman, Duco Veen, Joukje Willemsen, and Christopher Yau. 2021. Bayesian statistics and modelling. *Nature Reviews Methods Primers* 1, 1 (Jan. 2021), 1–26. https://doi.org/10.1038/s43586-020-00001-2
[77] Don van den Bergh, Francis Tuerlinckx, and Stijn Verdonck. 2020. DstarM: an R package for analyzing two-choice reaction time data with the D*M method. *Behavior Research Methods* 52 (May 2020), 521–543. https://doi.org/10.3758/s13428-019-01249-7
[78] Steeven Villa, Thomas Kosch, Felix Grelka, Albrecht Schmidt, and Robin Welsch. 2023. The placebo effect of human augmentation: Anticipating cognitive augmentation increases risk-taking behavior. *Computers in Human Behavior* 146 (Sept. 2023), 107787. https://doi.org/10.1016/j.chb.2023.107787
[79] Andreas Voss, Jochen Voss, and Veronika Lerche. 2015. Assessing cognitive processes with diffusion model analyses: a tutorial based on fast-dm-30. *Frontiers in Psychology* 6 (2015). https://doi.org/10.3389/fpsyg.2015.00336
[80] John R Wilson and Andrew Rutherford. 1989. Mental models: Theory and application in human factors. *Human Factors* 31, 6 (Dec. 1989), 617–634. https://doi.org/10.1177/001872088903100601
[81] Arkady Zgonnikov, David Abbink, and Gustav Markkula. 2022. Should I Stay or Should I Go? Cognitive Modeling of Left-Turn Gap Acceptance Decisions in Human Drivers. *Human Factors* (Dec. 2022), 15 pages. https://doi.org/10.1177/00187208221144561
[82] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, Cscw3 (Jan. 2021), 1–25. https://doi.org/10.1145/3432945

A VERBAL DESCRIPTIONS OF THE SYSTEM

The following is the explanation provided to the participants in the negative verbal description condition:

> *The first users of ADAPTIMIND^{TM} reported that when using the system, it decreased their task performance and increased stress making the task more difficult. As it is a new and untried AI system, it is very unreliable and risky to implement in real-world applications. In this study, we want to test these preliminary findings in a controlled setting. We would like to evaluate your performance using AI and compare it to a*

condition where the AI is inactive (control condition). We will remind you in which of the two conditions you are in before starting the tasks.

In the positive verbal description condition, only the system description was altered to reflect the system positively:

The first users of ADAPTIMIND^{TM} reported that when using the system, it increased their task performance and decreased stress, making the task easier. As it is a cutting-edge AI system, it is very reliable and safe to implement in real-world applications. In this study, we want to test these preliminary findings in a controlled setting.

B MAILS, TIA AND SIPAS

Table 6. Mean scores and standard deviation as a function of Description for the questionnaires Meta AI Literacy Scale (MAILS), Checklist for Trust between People and Automation (TiA) and Subjective Information Processing Awareness Scale (SIPAS)

	MAILS		TiA		SIPAS	
Description	*M*	*SD*	*M*	*SD*	*M*	*SD*
Negative	108.61	28.28	47.97	9.15	3.47	1.05
Positive	119.38	27.69	47.94	10.25	3.63	1.05

HIERARCHICAL DRIFT DIFFUSION MODEL WITH STATUS AND DESCRIPTION IN BRMS

All parameters are modeled on the log scale using the Wiener distribution.

(1) **Drift rate (v)**:

$$\log(v_{ijk}) = \beta_{0v} + \beta_{1v} \cdot \text{Status}_j + \beta_{2v} \cdot \text{Description}_k + \beta_{3v} \cdot \text{Status}_j \times \text{group}_k + b_{iv} \quad (1)$$

(2) **Boundary separation (α)**:

$$\log(\alpha_{ijk}) = \beta_{0\alpha} + \beta_{1\alpha} \cdot \text{Status}_j + \beta_{2\alpha} \cdot \text{Description}_k + \beta_{3\alpha} \cdot \text{Status}_j \times \text{Description}_k + b_{i\alpha} \quad (2)$$

(3) **Non-decision time (τ)**:

$$\log(\tau_{ijk}) = \beta_{0\tau} + \beta_{1\tau} \cdot \text{Status}_j + \beta_{2\tau} \cdot \text{Description}_k + \beta_{3\tau} \cdot \text{Status}_j \times \text{Description}_k + b_{i\tau} \quad (3)$$

Parameters and Priors:

- Intercept priors:

$$\beta_{0v} \sim \text{Normal}(0.74, 0.5)$$
$$\beta_{0\alpha} \sim \text{Normal}(0.40, 1),\ \text{lb} = 0.1$$
$$\beta_{0\tau} \sim \text{Normal}(-15, 1),\ \text{lb} = \text{-25},\ \text{ub} = 3$$

- Slope priors:

$$\beta_{1v}, \beta_{2v}, \beta_{3v} \sim \text{Normal}(0, 0.5)$$
$$\beta_{1\alpha}, \beta_{2\alpha}, \beta_{3\alpha} \sim \text{Normal}(0, 0.1)$$
$$\beta_{1\tau}, \beta_{2\tau}, \beta_{3\tau} \sim \text{Normal}(0, 0.01)$$

- Random effects:

$$b_{iv}, b_{i\alpha}, b_{i\tau} \sim \text{Normal}(0, \sigma)$$

Reaction Time Modeling:

$$f(RT|\log(\nu_{ijk}), \log(\alpha_{ijk}), \log(\tau_{ijk}), \text{bias} = 0.5) \tag{4}$$

Where RT is the observed reaction time.

C DEVIATIONS FROM THE PRE-REGISTRATION

Table 7. Explanations for deviating from the pre-registration

Section	Deviation
Participants: Recruiting and testing	We deviated from first testing 46 participants for nocebo (negative description), followed by testing 46 for placebo (positive description) due to time constraints. We stopped testing the negative description group after 30 participants were reached and then proceeded with testing the positive description group until we reached 60 participants. After this, we alternated the allocation of the last 6 participants to each group. The last day of testing remained the 18th of August 2023.
Reaction time data: Excluding trials	We excluded trials with too short responses by filtering RT under 150 ms instead of under 300 ms. This was a necessary deviation as participants were faster in their reactions than anticipated.
Reaction time data: Group Analyses	Given the AI performance bias, we modeled the data of both groups together instead of separately.

D MODEL PARAMETERS OF THE DDM

Table 8. Model outputs for the parameters on the log scale. Medians are provided for each parameter, along with their 95% HDI and p_b. Parameters distinguishable from zero are marked with *. We ran the model with two chains and 4000 iterations.

Parameter	Median	95% HDI	p_b
ν - Intercept	0.68	[0.55, 0.79]	0.00%
α - Intercept	0.37	[0.32, 0.42]	0.00%
τ - Intercept	-1.20	[-1.28, -1.13]	0.00%
ν Status*	0.04	[0.02, 0.05]	0.00%
ν Description	0.05	[-0.06, 0.17]	18.22%
ν Status × Description	0.01	[0.00, 0.02]	9.42%
α Status*	0.02	[0.01, 0.03]	0.03%
α Description	0.03	[-0.01, 0.07]	10.31%
α Status × Description	0.01	[-0.01, 0.01]	19.11%
τ Status*	-0.02	[-0.02, -0.01]	0.00%
τ Description	0.00	[-0.02, 0.02]	49.97%
τ Status × Description*	-0.01	[-0.01, 0.00]	0.12%