Int. J. Human–Computer Studies 198 (2025) 103474

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Machine learning systems as mentors in human learning: A user study on machine bias transmission in medical training

Lucia Vicente [a,*], Helena Matute [b], Caterina Fregosi [a], Federico Cabitza [a,c]

[a] Università degli Studi di Milano-Bicocca, Viale Sarca 336, Milano, 20126, Italy
[b] Universidad de Deusto, Unibertsitate Etorb., 24, Deusto, Bilbao, 48007, Spain
[c] IRCCS Ospedale Galeazzi-Sant'Ambrogio, Via Belgioioso 173, Milano, 20157, Italy

ARTICLE INFO

Dataset link: https://osf.io/uxy79/

Keywords:
Bias transmission
Human-AI collaboration
Hybrid intelligence
Machine mentoring
Learning

ABSTRACT

While accurate AI systems can enhance human performance, exerting both an augmentation and good mentoring effect, imperfect systems may act as poor mentors, transmitting biases and systematic errors to users. However, there is still limited research on the potential for AI to transmit biases to humans, an effect that could be even more pronounced for less experienced users, such as novices or trainees, making decisions supported by AI-based systems. To investigate the bias transmission effect and the potential of AI to serve as a mentor, we involved eighty-six medical students, dividing them into an AI-assisted group and a control group. We tasked them with classifying simulated tissue samples for a fictitious disease. In the first phase of the task, the AI group received diagnostic advice from a simulated AI system that made systematic errors for a specific type of case, while being accurate for all other types. The control group did not receive any assistance. In the second phase, participants in both groups classified new tissue samples, including ambiguous cases, without any support to test the residual impact of AI bias. The results showed that the AI-assisted group exhibited a higher error rate when classifying cases where the AI provided systematically erroneous advice, both in the AI-assisted and the subsequent unassisted phase, suggesting the persistence of AI-induced bias. Our study emphasizes the need for careful implementation and continuous evaluation of AI systems in education and training to mitigate potential negative impacts on trainee learning outcomes.

1. Introduction

The impact of artificial intelligence (AI) on decision-making performance has been extensively studied (Lai et al., 2023). Numerous studies converge to conclude that this effect is generally positive, often referred to as augmentation (Fügener et al., 2022; Castaneda et al., 2015). However, some studies have also observed a degradation in performance due to automation bias, that is machine-induced errors that occur when decision makers fail to recognize incorrect suggestions (Parasuraman and Manzey, 2010; Vaccaro et al., 2024).

One area that has received relatively little attention is the long-term effect of AI use. Some observers have highlighted the risk of skill erosion (deskilling) among experts (Cabitza et al., 2017) and a negative impact on staff in training (Beane, 2019), particularly concerning their skill acquisition (upskilling inhibition) and autonomy. It is therefore interesting to complement existing research on the immediate effects of AI-assisted decision-making with research on the effects of AI on subsequent *unaided* decision-making. Doing so would help understand the role of AI as a *mentor*, guiding by example and transmitting optimal decision-making practices, rather than serving as an indispensable assistant necessary for the human decision maker to attain a certain level of performance.

The notion that competent (i.e., highly accurate) machines can effectively "train" humans in interaction and "by doing", and lead them to better learning outcomes than they would achieve on their own or with the help of experts, has its origins in early human-AI interaction settings such as chess (Gaessler and Piezunka, 2023) and in the seminal ideas of B.F. Skinner, the famous psychologist who revolutionized the psychology of learning in the late fifties (Skinner, 1958, 1961). The increased availability of machines, their competence comparable to that of experts in limited tasks, and their responsiveness are some of the reasons for the potential of machines in human training. However, this list is not exhaustive and does not include what may be the most important factors in this phenomenon.

In addition, machine *training in interaction*, or *machine mentoring*, often involves two additional and powerful factors of human cognition, distinguishing it from cases where machines are intentionally used in

* Corresponding author.
E-mail address: lucia.vicenteholgado@unimib.it (L. Vicente).

https://doi.org/10.1016/j.ijhcs.2025.103474
Received 16 August 2024; Received in revised form 7 February 2025; Accepted 11 February 2025
Available online 20 February 2025

education and training. First, *ethopoiesis* (also known as *ethopoeia*) (Nass and Moon, 2000), which is the human innate ability to project human capabilities and characteristics onto inanimate objects, particularly computational agents, and consider them *social actors*. Second, the propensity of humans for *cultural learning* (Heyes, 2017), which is the tendency of learners and trainees to quickly adopt the behaviors of those social actors they deem to be *models* (Bandura and Walters, 1977), i.e., actors associated with a reputation of excellent performers and experts, i.e. what is called *prestige* (Henrich and Gil-White, 2001).

A recent example of cultural learning involving artificial intelligence as a model of behavior can be found in the work of Shin et al. (2023), where authors analyzed the historical record of Go players' strategies before and after the advent of AlphaGo as a potential opponent. This research provided evidence of how interacting with an AI capable of exhibiting excellent performance resulted in significant changes in autonomous human decision-making, leading to higher accuracy and creativity as if users had learnt taking inspiration and emulating the AI behavior.

In this paper, we aim to explore this phenomenon: whether humans *internalize* and acquire patterns and decision criteria (including implicit and subsymbolic ones) merely through interaction with machines, even when these machines do not have the specific function of a tutor. In other words, we aim to verify whether machine-human knowledge transmission occurs, not only in the traditional paradigm where humans give their past best decisions to train machines (the machine learning approach), but also in the opposite direction (Chen, 2024), where machine decisions and advice can shape humans' behavior: the machine training-on-the-job approach, especially when humans perceive themselves as less proficient than the machine, and regard it as a model within a social learning context.

The starting hypothesis of our study is then that AI systems, having been trained on extensive and high-quality knowledge sources produced by humanity over the past centuries — such as encyclopedias, textbooks, manuals, academic articles, and non-fictional books — along with thousands of vetted and well-documented decisions (e.g., supervised medical decision support systems), can "act as models of human culture, facilitating cultural transmission across individuals and generations" (Brinkmann et al., 2023), especially when learners are involved.

The AI scientific community is obviously aware that, despite their prestige, these systems are not, and cannot be, perfect classifiers, and therefore can be *bad models*, even without intending to. This imperfection arises either from errors in the data used during their development (Cabitza et al., 2019) or from their failure to extrapolate correct answers when applied to new cases.

In this paper, we focus on this phenomenon of cultural and social learning by humans when their models are machines, occurring in a hybrid human-machine socialization process (Nonaka and Takeuchi, 2007) to understand the extent to which these model systems are capable of transmitting knowledge, both accurate and biased, as this effect can be assessed by observing the performance of trainees when they are not using these systems.

Bias transmission is distinct from automation bias or *induced belief revision* (Kwong et al., 2024; Goddard et al., 2012): while several studies have focused on how AI can systematically bias decision-makers with its advice (Adam et al., 2022; Agudo et al., 2024; Dratsch et al., 2023; Kupfer et al., 2023), thereby objectifying, reiterating, and perpetuating the biases inherent in its training data, only one study to our knowledge has addressed the case of humans learning biased decision criteria and applying them when unassisted, referred to as *bias inheritance* (Vicente and Matute, 2023).[1]

This study aims to fill this research gap by exploring how low-expertise users, such as trainees, students and novices, can unconsciously adopt poor pattern matching heuristics or misleading decision criteria from the machines while making decisions with their support and subsequently replicate these criteria *when not using the machines*.

While in Vicente and Matute (2023), bias transmission was studied in a general population sample, we studied this effect in medical students working with an AI in a simulated medical setting. When students and trainees use a supposedly accurate diagnostic support system, they may learn not only correct decision patterns, but also incorrect ones, which we here refer to simply as *biases*,[2] that is when AI systems come up being bad trainers.

2. Methods

2.1. Ethics statement

The Ethical Review Board of the University of Deusto (Bilbao, Spain) reviewed and approved the methodology described in this article. The experiment was conducted according to the approved protocol. Informed consent was obtained from all volunteers prior to participation, and no personal or sensitive information was collected. For ethical reasons and to prevent prior knowledge or belief from influencing the results, all elements of the experimental setting, including the clinical context, the classification task, the artificial intelligence system, the images of the tissue samples and the syndrome, were fictitious.

2.2. Participants

In our study we involved a total of 86 volunteers (78% female, 21% male, 1% preferred not to answer) with an average age of 19.8 years (SD = 1.62; minimum = 18, maximum = 28). There was no need to exclude data from any participant according to the data selection criteria described in the Materials and Procedure section. The sample included first-year (43%) and third-year (57%) students of Medicine. Participants were randomly assigned to either the AI group (n = 43), which performed a diagnostic task supported by a fictitious AI-based decision support system (DSS), or the control group (n = 43), which completed the same tasks without assistance.

2.3. Materials and procedure

The experimental task was carried out through Qualtrics, an online questionnaire platform. To simulate a clinical diagnostic task, we used cases created by Blanco et al. (2020). Each case consisted of a 50 × 50 pixel matrix representing a human tissue sample. Each tissue sample contained 2500 cells of two colors (dark pink and light yellow) randomly distributed in the matrix to ensure unique samples. The proportion of dark and light cells varied, resulting in a diverse set of cases with varying levels of diagnostic difficulty. See Table 1 for different proportions of dark and light cells presented to participants.

In the experiment, participants were asked to observe a series of tissue samples from different patients to determine whether each

[1] The terms transmission and inheritance are both evidently metaphoric in nature. The term transmission aligns more closely with (and is inspired by) the populationist approach to cultural evolution theory, often associated with the Californian School perspective (Henrich, 2016). In contrast, inheritance is more commonly linked with the selectionist or Darwinian approach (Heyes, 2018) to that theory. We favor the former term, adopted also in Brinkmann et al. (2023), emphasizing that transmission should not be understood in the Shannonian sense of token dispatch or faithful duplication. Instead, it should be interpreted like in the context of disease transmission, where pathogens (ie ideas, beliefs, behaviors, practices) can mutate and adapt to a new host (i.e., the learner) after being socially transmitted by a teacher or cultural model.

[2] we define bias without any ethical connotation, but only, as is usually used in psychology, as a systematic and recurrent error (Kahneman, 2011), that is with an emphasis on learnability.

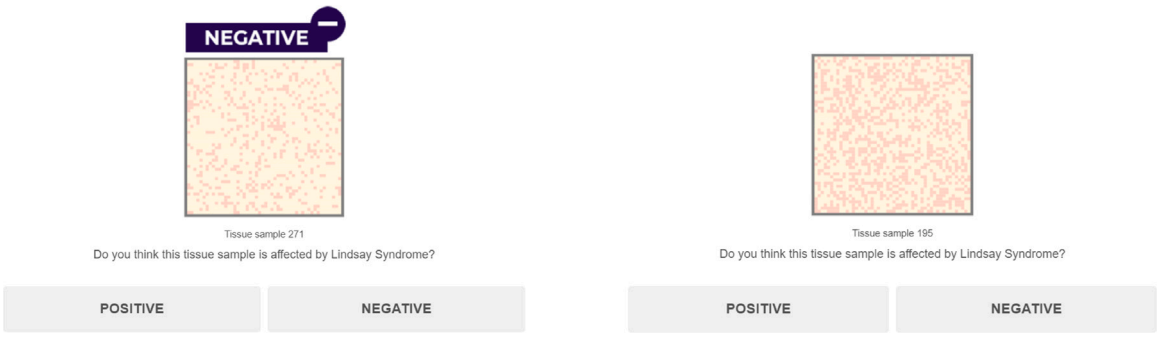

Fig. 1. (a) (left): The image displays a screenshot of the classification task presented to participants with AI support, indicated by the 'NEGATIVE' label at the top. (b) (right): The image shows a screenshot of the classification task given to participants in the control group, without AI support.

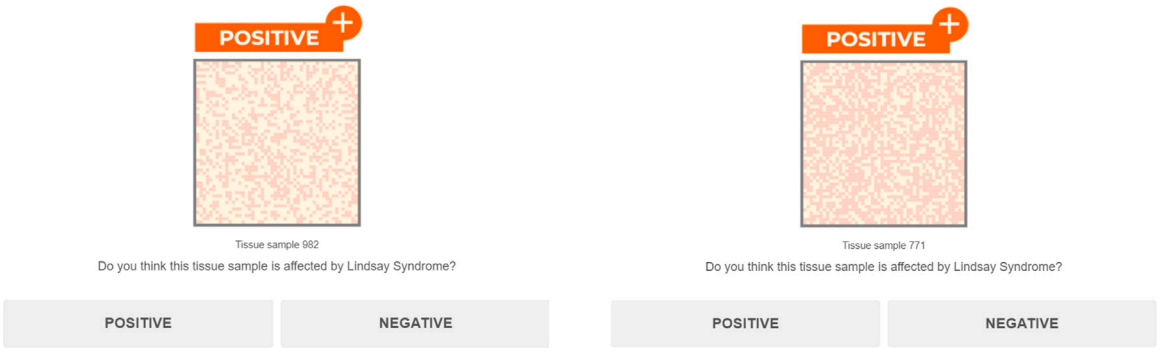

Fig. 2. (a) (left): The image shows a 40/60 Hard Case with the systematic error presented by the AI, indicated by the "POSITIVE" label at the top. (b) (right): The image depicts a 60/40 Hard Case with the correct classification made by the AI.

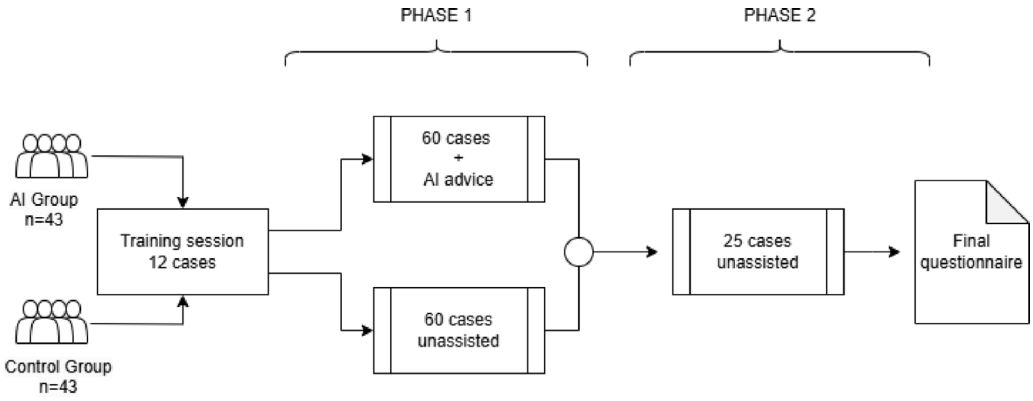

Fig. 3. This diagram illustrates the study design and participant distribution. Both the AI Group (n = 43) and the Control Group (n = 43) received the same training, in which they had to correctly rate 12 cases: 4 hard cases and 8 easier cases presented in random order. Then they were involved in the 1st phase, where both groups were required to rate 60 cases each: 20 hard cases and 40 easier cases, presented in random order. In this phase, the AI Group received AI advice during the classification of all these cases, while the Control Group did not. Subsequently, in the 2nd phase, both groups completed 25 additional unassisted cases: 10 hard cases, 5 ambiguous cases, and 10 easier cases presented in random order. Finally, all participants filled out a post-test questionnaire.

Table 1
Case types based on the proportions of dark and light cells (experimental stimuli). We labeled the 40/60 and 60/40 cases as hard cases because they were more difficult to discriminate than cases with clearer ratios, such as 80/20 or 30/70. Thus, "hard" and "easier" are relative terms. These labels are only intended to help distinguish the different types of cases.

Cell proportion	Case type
40/60, 60/40	Hard cases
20/80, 30/70, 70/30, 80/20	Easier cases
50/50	Ambiguous cases

sample was affected or not by a fictitious disease known as Lindsay syndrome. Each sample consisted of cells in two colors, with one color predominating in all samples except that were hence denoted as ambiguous samples. Participants were instructed to observe the proportions of the two colors and use a simple diagnostic rule based on the predominant color to determine the presence of the syndrome. Which of the two colors determined the presence, or the absence, of the syndrome was randomized for each participant. In one version of the instructions (Version 1), a higher proportion of dark-pink cells than light-yellow cells indicated that the tissue sample was affected by the fictitious syndrome and was "Positive" to the condition, whereas a greater proportion of light-yellow cells than dark-pink cells indicated that the samples were not affected by the syndrome and should be categorized as "Negative". Conversely, in the other version of the instructions (Version 2), a higher proportion of light-yellow cells in

the tissue sample indicated a positive, while a greater proportion of dark-pink cells signified a negative.

The experiment began with a training phase (see Fig. 3) that consisted of two blocks of six cases presented one per trial, that is, 12 trials in total. In the first block, the six samples were presented in order of difficulty (see Table 1). In the second block, the same six samples were presented in random order. Participants who did not achieve at least five correct classifications out of six trials had to reread the instructions and repeat the entire training phase. If they did not get at least 5 correct answers out of 6 in this second repetition of the training phase, the data from these participants were excluded from the analysis. Once this phase had been completed, participants were randomly assigned to the AI group and control group and begun the Phase 1 of the experiment.

2.3.1. Phase 1

At the beginning of Phase 1, all participants read the same description of the AI system:

> Observing tissue samples can help us diagnose Lindsay Syndrome at an early stage, improving patients' chances for treatment and recovery. Recently, a diagnostic assistance system based on an artificial intelligence (AI) algorithm has been developed to help doctors in the early detection of Lindsay Syndrome. This system analyses each tissue sample and provides recommendations during its classification.

On a second screen, the instructions followed as described here:

> We have received many tissue samples from a group of people. Some are suffering from Lindsay Syndrome, while others are not. We are asking you to help us classify these samples. During the classification task, the artificial intelligence system may be connected. If so, it will offer you its recommendation. If the AI is on it will offer you a recommendation that will appear at the top of the screen above the sample, as you can see in the image. But if the AI is off, you will have to perform the whole task without assistance.

This description avoids providing information about the accuracy or any other characteristic of the simulated AI system. We aimed to prevent creating any expectations about the model in the participants. Specifically, we wanted to avoid fostering either excessive trust or distrust in the model, as this would have influenced participants' decision to follow the AI's advice.

Phase 1 of the experiment involved 60 tissue samples. If participants did not perform better than chance (more than 30 out of 60 hits) during Phase 1, which would indicate that they did not pay attention to the task or did not understand the instructions, their data were excluded from the analysis. The fictitious AI of our experiment was "switched off" for the control group, so they had to classify these tissue samples in Phase 1 without any assistance. Conversely, in the AI group, the AI offered its recommendations during Phase 1. Thus, in the AI-assisted group each trial involved presenting the tissue sample for classification along with the AI recommendation for that case, in a so-called concurrent (Tejeda et al., 2022) or AI-first (Cabitza et al., 2023b) interaction protocol. The AI advice appeared as an orange label with the text "POSITIVE+" or a blue label with the text "NEGATIVE-" placed over the image of the tissue sample (see Fig. 1). We specify that our protocol was concurrent, which has been a common approach for presenting AI advice in some real-world contexts, to distinguish it from human-first protocols (Cabitza et al., 2023b) or update settings (Green and Chen, 2019) where humans make an initial decision, next view AI advice and then they revise their decision accordingly.

The sequence of trials in the classification task included ten cases of each of the hard cases and easier cases proportions as shown in Table 1, for a total of 60 cases. Half of those cases were positive, and half were negative, following the correct classification criteria. The simulated AI provided correct recommendations for 50 out of 60 tissue samples (accuracy 83%), but systematically failed in its recommendations for the ten 40/60 cases (see Fig. 2). In version 1 of the instructions, the AI recommended classifying these cases as positive when the correct classification would have been negative. In version 2, the AI recommended to classify the 40/60 tissue samples as negative when they should have been considered positive.

The order of tissue samples was randomly assigned across the 60-trial sequence with one exception. The 40/60 hard cases, where the AI advice was always incorrect, were not included in the first 10 trials of the sequence, not to foster a negative halo effect or distrust of the AI (Tractinsky et al., 2000; Vered et al., 2023; Yu et al., 2019).

In no case did participants receive trial-by-trial feedback after their responses, nor any information about their accuracy or error rates in the task. After completing this first phase, all participants could begin the second phase (see Fig. 3).

2.3.2. Phase 2

During Phase 2, both groups classified the tissue samples without any assistance. We included five cases of each cell proportion, except for cases with dark/light cell ratios of 80/20 and 20/80, which were excluded from this phase because they were considered too easy. In addition, Phase 2 included five new trials with ambiguous cases having a 50/50 ratio of dark/light cells. The classification criterion, as outlined in the initial instructions, was based on the predominant color of the tissue samples for assigning them to positive or negative categories. Since the 50/50 samples had no predominant color, it was impossible to assign these cases to either category based on the objective information in the tissue sample. This resulted in a total of 25 unassisted trials. According to correct classification criteria ten of them were positive, ten of them were negative, and five were ambiguous. The most interesting aspect of this second phase was recording participants' responses to the 40/60 cases and the 50/50 ambiguous cases, to see if they had acquired (learned) any AI bias.

The order of appearance for each type of case in Phase 2 was determined by a fixed, random sequence. Both participants that in Phase 1 had been assisted by the AI, and those in the unassisted group observed the same sequence of trials. As in Phase 1, participants in Phase 2 did not receive any feedback on their performance in the classification task.

2.3.3. Post-test questionnaire

After Phase 2, participants completed a post-test questionnaire to report their performance and perceptions of the AI's advice during the classification task. The questionnaire included the following items, rated on a 9-value ordinal scale with options ranging from 1 (not at all) to 9 (completely):

QH (Helpfulness): "To what extent do you think the AI's algorithm has been helpful during the tissue sample classification task?"

QR (Reliance): "To what extent have you followed the AI's recommendations during the classification task?"

QA (Accuracy): "How accurate was the artificial intelligence's algorithm in classifying the tissue samples?"

QT (Trustworthiness): "To what point do you consider trustworthy the advice that an artificial intelligence can offer, in general, in the field of health?"

QE (Error detection): "Have you detected any errors in the AI's recommendations?" Yes/No answer.

The main aim of our study was to test whether the bias shown by the AI in Phase 1 (i.e., the systematic erroneous suggestion for 40/60 samples), would affect participants' response tendencies in the subsequent unaided Phase 2, particularly for the 40/60 cases and the ambiguous 50/50 cases. Similarly, we wanted to test whether any improvements in participants' performance during Phase 1, in cases where the AI provided accurate recommendations, could be maintained when the support of the AI was removed. However, before analyzing the bias transmission and machine mentoring effects in Phase 2, we

Table 2
Results of the two-sample proportion test comparing the error rates between the AI assisted and control group across Phase 1 and Phase 2 for different case types. Values are provided alongside their corresponding p-values, which indicate the statistical significance of the differences observed, and effect sizes, measured using Cohen's h, which quantifies the magnitude of the difference in proportion.

	Error rates AI vs. Control	P-value	Effect size
40/60 cases Phase 1	0.35, 0.01	< .001	= 1.03
40/60 cases Phase 2	0.26, 0.01	< .001	= 0.88
All other cases Phase 1	0.00, 0.01	= .003	= 0.10
All other cases Phase 2	0.00, 0.02	= .012	= 0.15
Ambiguous cases Phase 2	0.76, 0.39	< .001	= 0.78

first needed to test whether participants relied on the AI's help in Phase 1, both in cases where the AI gave accurate advice and in those cases where the AI was wrong. In other words, we wanted to test whether accurate AI assistance would improve the accuracy of the AI-assisted participants in Phase 1, thereby demonstrating the AI-induced augmentation effect previously documented in the literature. Similarly, we examined whether participants would accept and follow incorrect AI suggestions, replicating the well-documented automation bias effect. We conjectured that reliance on the AI's advice would be necessary for the model to exert a sustainable influence on people's behavior. This reliance may lead individuals to imitate the machine and internalize its response patterns, whether they intend to learn from it or not. The post-test questionnaire would complement this behavioral assessment of participants' reliance, providing qualitative insights into their self-reported trust and self-perceived reliance on artificial intelligence.

3. Results

To address our main research questions, we compared error rates between the AI group and the control group across Phase 1 and Phase 2 for different types of cases (see Table 1). We utilized the Two-Sample Proportion Test for this comparison. Detailed test results, including p-values and effect sizes, are presented in Table 2, with summarized outcomes graphically depicted in Fig. 4.

In Phase 1, we observed a significant difference in the ratio of classification errors in 40/60 cases, those presented to the AI group along with the incorrect AI advice, between the AI-assisted group and the control group ($p < .001$), as detailed in Table 2. This difference, illustrated by the dashed line labeled AB in Fig. 4, suggests that participants relied on the systemically erroneous AI recommendations for the 40/60 cases.

For all other cases, excluding those with the 40/60 proportion, represented by the dashed line labeled AE in Fig. 4, the two-sample proportion z-test indicated a small yet significant difference in error rates between the AI and control groups ($p = .003$) in Phase 1. Participants assisted by the AI demonstrated the lowest error rates (see Table 2). This finding suggests a positive effect of AI support in the classification task, although the improvement was modest in this case. The control group's error rates were quite low, indicating that the task was easy to perform and that there was limited potential for further improvement with AI assistance.

In summary, we found that the AI's incorrect suggestions led to errors among AI-assisted participants during Phase 1. Our results support the automation bias effect: participants seemed to rely too heavily on the AI's recommendations, which influenced their responses even in a task that they could have performed effectively without assistance, as shown by the lower error rates of the unassisted group. This also means that AI errors might have been relatively easy to identify, but even so, systematic errors influenced participants' responses. Conversely, when the AI provided accurate support, it slightly improved the performance of assisted participants in Phase 1.

In Phase 2, participants in both groups classified tissue samples without AI assistance. This means that those in the AI-assisted group, who had support in Phase 1, were required to classify samples without any kind of assistance in Phase 2. We found evidence of bias transmission after interaction with a biased AI system. Participants who relied on AI advice in Phase 1 continued to make the same type of systematic errors as the model in Phase 2. The two-portion test (see Table 2) revealed that the AI group still made significantly more mistakes than the control group in the 40/60 cases, even when the misleading AI assistance was removed ($p < .001$). The dashed line labeled BTM in Fig. 4 illustrates the significant and large effect observed in Phase 2, highlighting the bias acquired during the assisted classification in Phase 1 for the 40/60 cases.

We recorded participants' tendency to classify ambiguous cases (i.e., those with a 50/50 dark/light cell proportion) in the same category as suggested by the incorrect AI recommendations for 40/60 cases in Phase 1. The results of the two-sample proportion test for "biased" classification of ambiguous cases revealed a significant medium-sized difference ($p < .001$) between the AI group and the control group. This result, represented by the dashed line labeled BTF in Fig. 4, indicates that the bias learned from the AI for the 40/60 cases in Phase 1, impacted the participants' classification of ambiguous 50/50 cases in Phase 2. We interpreted this finding as a generalization of the AI-induced bias.

Finally, the dashed line labeled UE indicates that the AI-assisted group had significantly lower overall error rates compared to the control group for all cases (excluding 40/60 cases) in Phase 2 ($p = .012$). This finding suggests a positive and significant, albeit small, persistent impact of AI advice, even after it was no longer directly available. Therefore, the performance improvement gained from interacting with accurate AI assistance seemed to be maintained even when that support was removed, indicating a modest *upskilling* or *machine training effect*.

Additionally, we analyzed the *technology impact* (Cabitza et al., 2023a) on different types of cases in both Phase 1 and Phase 2, with results illustrated in Fig. 5.

The metric known as technology impact represents the ratio of the likelihood of an error in the AI group compared to the likelihood of an error in the control group. When technology impact values are larger than 1 an overall positive effect has occurred, and the AI intervention had a useful effect. Conversely, values lower than 1 denote a detrimental effect of AI on decision-making. As shown in Fig. 5, in Phase 1, the impact of the AI support on cases where this was totally accurate was significantly positive, confirming the slight but significant difference with respect to the control group, represented by the AE dashed line (see Fig. 4). In Phase 2, despite the removal of AI support, we could observe a residual and persistent positive impact, especially for the type of cases that in Phase 1 had not been associated with inaccurate advice by the AI. Conversely, for 40/60 cases and ambiguous cases, the technology impact is negative in both phases, as indicated in Fig. 5. This could be a reflection of the misleading classification by AI that had been acquired by the subjects, or so to say transmitted by the AI, which produces a persistent negative impact even when the AI support was no longer present in Phase 2. Interestingly, the technology impact for all other cases is similar across both phases, suggesting a persistent effect of the initial AI support, when this is accurate and hence reliable.

Regarding the post-test questionnaire, we performed a Spearman rank correlation analysis to examine the relationship between participants' error rates in 40/60 cases and their responses to each of the questions. This analysis was conducted separately for Phase 1 (see Table 3) and Phase 2 (see Table 4) of the classification task. The worse the impact of AI on decision-making (in terms of accuracy in Phase 1 and 2), and therefore the higher the automation bias observed, the more helpful, reliable, accurate and even trustworthy the decision-makers perceived the AI, showing a significant and strong correlation

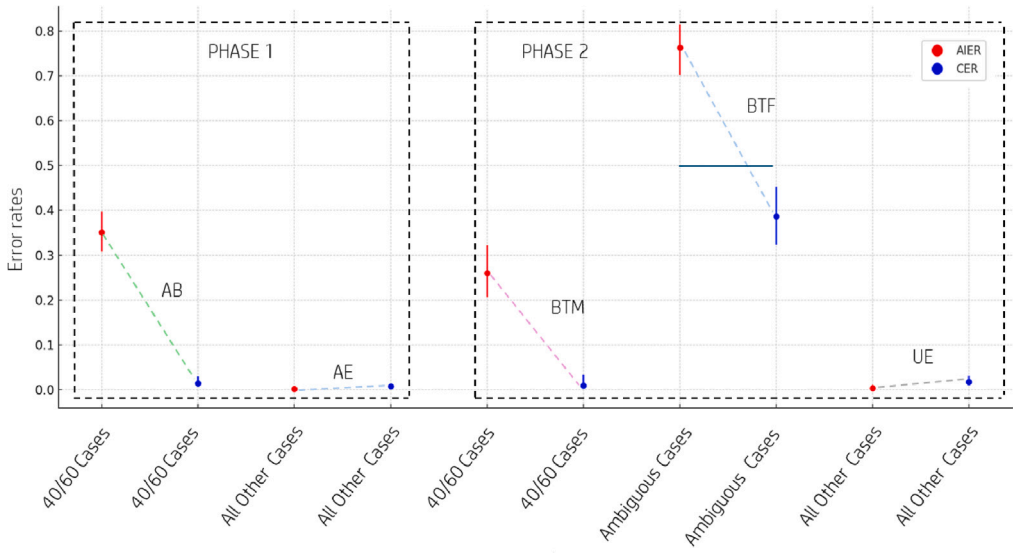

Fig. 4. Error rates and their 95% confidence intervals are compared between the AI-supported group (AIER) and the control (CER) group during Phase 1 (on the left) and 2 (on the right). Dashed lines are drawn to facilitate meaningful comparisons, which must be made for same-type cases. Comparison AB (Automation Bias) highlights the effect of automation bias, which refers to the errors made by participants who relied on AI that only provided incorrect answers. Dashed line AE (Augmentation Effect) indicates a marginal positive impact of AI, as the AI group exhibited a slightly lower error rate compared to the control group in Phase 1, except for hard cases. Dashed line BTM (Bias Transmission) demonstrates the effect of acquired bias, suggesting that previous reliance on AI influenced participants' judgment on similar hard cases in Phase 2, even without AI support. Dashed line BTF (Bias Generalization) regards the extension of this acquired bias to ambiguous cases that were not shown in Phase 1. Dashed line UE (Upskilling Effect) shows that the AI group had lower error rates in Phase 2 for simpler cases, implying a positive residual impact of AI technology as an effective mentoring tool. The comparison denoted by BTF does not actually regard error rates, but the frequency with which respondents chose the label associated with a wrong answer in Phase 1, and this is why the 50% threshold of random choice is also highlighted.

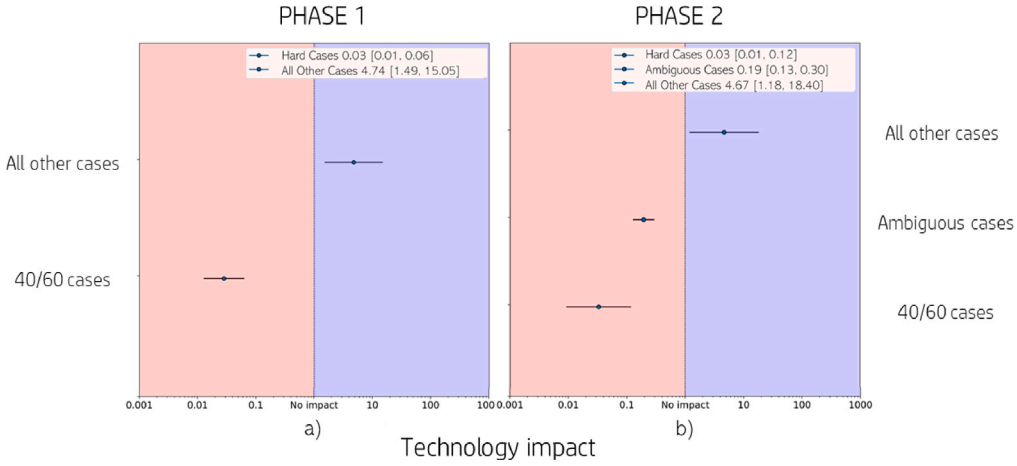

Fig. 5. The diagram illustrates the Technology Impact across various case types and experimental phases, computed by comparing the performance of the AI-supported group with the control (unaided) group. The diagram uses a logarithmic scale to display odds ratios, where the red region (on the left) indicates an overall negative effect of the AI support (which only provided incorrect answers for the hard cases), while the blue region (on the right) indicates an overall positive effect in terms of accuracy improvement and error rate reduction. In (a) (on the left), figures represent the impact of technology on AI-supported users. In (b) (on the right), the effect of AI refers to the impact of having received the AI support during the previous learning phase (phase 1), therefore showing a persistent impact of technology. Generated with the tool available online at https://mudilab.github.io/dss-quality-assessment/.

Table 3
Spearman's rank correlation scores between error rates and the psychometric constructs in Phase 1 reported by the AI users.

Construct	Correlation	P-value
QH: AI system perceived helpfulness	.74	< .001
QR: Reliance on AI recommendations	.71	< .001
QA: AI system perceived accuracy	.51	.003
QT: AI system perceived trustworthiness	.51	< .001

Table 4
Spearman's rank correlation scores between error rates and the psychometric constructs in Phase 2 as perceived by the participants who used the AI system in Phase 1.

Construct	Correlation	P-value
QH: AI system perceived helpfulness	.63	< .001
QR: Reliance on AI recommendations	.63	< .001
QA: AI system perceived accuracy	.61	< .001
QT: AI system perceived trustworthiness	.47	= .002

between error rates and the constructs explored in the questionnaire. This can be interpreted as a sign of that a positive attitude towards AI increase the proneness to over-reliance.

The results of the Mann–Whitney U test, with a *p*-value of 1, indicate no statistically significant difference in responses to QT (regarding trust in AI advice) between the AI group and the control group. This finding

suggests that participants who used an AI with systematic errors, and continued to make those errors even without AI support, did not experience a loss of trust compared to the control group.

4. Discussion

Artificial intelligence systems that outperform humans are gradually being introduced in a wide variety of domains. Such systems, in their advisory role, can enhance human performance, with improvements that may persist even after assistance is withdrawn. The results of our study suggest that humans can learn from the AI systems they interact with. While this interaction can have a positive impact, as observed in other studies (Shin et al., 2023), AI systems can also act as "bad mentors", transmitting biases to human learners. In particular, *bias transmission* is a distinct and persistent effect compared to the more extensively studied phenomenon of *automation bias*. While automation bias concerns the potential misleading effects of AI-based Decision Support Systems (DSSs), especially in complex cases (Cabitza et al., 2023b), *bias transmission* exerts its influence even when decision makers do not use the system, and it affects not only cases where the machine provided incorrect suggestions but also different, yet similar, cases; a phenomenon, that we propose to distinguish from the former one, and call *bias generalization*.

Regarding the augmentation effect of AI on human decisions, which has been reported in several studies on AI-supported diagnosis (Fügener et al., 2022; Castaneda et al., 2015), we found a moderate positive impact of the AI support in the classification task. In Phase 1, AI support improved participants' performance in cases where this was accurate, confirmed by the slight but significant difference with the control group. This suggests that AI support was beneficial, possibly helping subjects avoid slips and lapses (Reason, 1990). In Phase 2, despite the removal of AI support, we could observe a residual and persistent positive impact, especially for the type of cases that had not been associated with inaccurate AI advice in Phase 1. Note that the effect size of the differences in error rates for cases with accurate AI advice between the AI-assisted and unassisted participants was small in both phases of the task. The control group performed the task quite well, and the introduction of AI support did not yield a substantial improvement, likely because there was limited room for larger effects.

However, our results also showed that excessive reliance on the machine could have a perdurable negative impact on human decisions: participants in the AI group relied on incorrect AI advice in Phase 1 (Goddard et al., 2012) and later they replicated the same mistakes as the machine in Phase 2, when classifying hard and ambiguous cases without any support. Thus, the bias acquired through interaction with AI manifested in the future unaided decisions of the participants and generalized to similar ambiguous cases. Although research on this phenomenon remains limited, evidence of bias transmission has been documented in previous studies using different experimental tasks and population samples (Vicente and Matute, 2023; Glickman and Sharot, 2023, 2024). The convergence of findings between studies suggests that bias transmission may be a robust effect.

Therefore, the potential transfer of knowledge from AI to humans is not without risk: in the case of imperfect or biased DSSs, humans can learn, generalize to new contexts and situations, and reproduce defective decision-making strategies, as outlined above. Regarding potential mechanisms involved in knowledge transfer from AI to humans, it has been suggested that access to AI reasoning processes and explanation provision might be necessary conditions for learning from AI, instead of relying on the mere observation of AI actions or responses (Shin et al., 2023). Nevertheless, repeated exposure to certain AI response patterns or repeated observation of case-label pairings seemed sufficient for the acquisition of a new classification criterion learned from the AI in our experiment besides the criteria shared during the training session that occurred before Phases 1 and 2. Humans are skilled pattern detectors (Holzinger et al., 2023) and natural-born social learners (Henrich, 2016): thus, they can easily infer rules after the observation of the actions of other agents, especially when these are considered competent (Gero et al., 2020). It appears that the participants were able to grasp the underlying rule established by the AI, internalize it, and subsequently incorporate it into their future decision-making.

The potential for AI systems to act as poor mentors by introducing biases in users' response patterns could be exacerbated when systems fail in complex cases where systematic errors may not be easy to detect. This risk becomes particularly pronounced for inexperienced users, who need more support and may not be sufficiently trained to be aware of machine errors. Although preliminary, given that our findings are based on a single experiment, our results suggest that a cautious and conservative approach is necessary when implementing oracular DDSs (that is decision support systems providing only advice without additional explanatory information) particularly for inexperienced or still-training individuals due to their susceptibility to acquiring such biases and even extrapolate them to similar cases.

Therefore, novices in training may be more susceptible to the negative effects of AI use, such as the medical students who participated in our research. The experimental task, adapted from Vicente and Matute (2023), simulated a medical diagnosis scenario. We wanted to determine whether the findings observed in a general population sample by these authors would be replicated in medical students, who might interpret elements of the experimental setting differently. Medical students may have experienced a greater sense of commitment and responsibility toward a task resembling those within their professional domain, even though the simulated diagnosis did not require specialized expertise. As a result, they could have been more cautious in the classification task and avoided over-reliance on the AI recommendations. Our findings replicated those of Vicente and Matute (2023): medical students' behavior in a clinical-related task was also susceptible to the influence of biased recommendations from an entity defined as artificial intelligence.

Future medical professionals will inevitably interact with artificial intelligence in their daily practice. Moreover, the current scenarios for future medicine presented to medical students place artificial intelligence at the forefront, envisioning optimistic possibilities for this technology (Rajpurkar et al., 2022; Topol, 2019). In addition, AI is already widely integrated into medical education and training further reinforcing its perceived credibility among students (Tozsin et al., 2024). For these reasons, medical students represent a particularly interesting population for examining the impact of systematic AI errors in a simulated medical decision-making task. While we acknowledge that the simulated nature of our experimental setup may limit the transferability of our results, we believe that our study illustrates the potential risk of bias transmission when imperfect AI systems are used in medical training.

Our study documented a bias transmission and machine mentoring effect from AI to humans in an experimental setting. Ours is one of the first works exploring these effects in an empirical lab experiment with human subjects. This fact represents both a strength and a limitation.

The strength of our study is that we report a novel finding with significant implications in several domains. Furthermore, we documented the effect in a controlled experiment, contributing to the growing research on human-AI decision-making. In recent years, the research community has increasingly recognized the need for empirical studies that examine how individuals interact with AI in decision-making tasks (Lai et al., 2023). Such studies are essential for building a foundational understanding of AI-assisted decision processes and their potential risks.

One of the main potential limitations of our research would be transferability, as we introduced earlier. Research on bias transmission and machine mentoring is limited, and the replicability and transferability of these effects have not been widely demonstrated. In addition,

the classification task used in our experiment can be perceived as a simplistic simulation of a complex medical decision-making process. Our task was indeed simple and participants could have chosen not to rely on the AI's suggestions. Similarly, they could have detected AI errors and avoided them. But they did not, or at least many participants were influenced by the AI's erroneous suggestions. We argue that if a flawed AI influenced performance in a task as simple as the one used in our experiment, the potential impact of AI errors could be even more pronounced in complex decision-making tasks. Such tasks require consideration of multiple factors and often involve greater ambiguity and uncertainty in interpreting information. In these contexts, the risk of AI misguidance could have even more detrimental effects on human decision-making.

In fact, the negative impact of erroneous AI recommendations has already been observed in complex clinical tasks, where AI-generated information led individuals to make mistakes they might have avoided had they relied on their own judgment (Adam et al., 2022; Dratsch et al., 2023; Jacobs et al., 2021; Rezazade Mehrizi et al., 2023; Gaube et al., 2021). For example, Jacobs et al. (2021) found that in a task that required prescribing antidepressants in different scenarios, incorrect recommendations by an artificial intelligence reduced the accuracy of specialists' clinical decisions compared to an unassisted control condition. Similarly, Adam et al. (2022) examined the influence of biased AI recommendations on expert responses to mental health emergencies. Their results showed that experts exhibited significant bias when advised by a biased AI, whereas no bias was observed when they made decisions without algorithmic support.

These findings demonstrate the impact of AI errors and biases on settings closer to real-world decision-making. Our study extends this research by showing that AI suggestions not only influence immediate aided decisions but also shape future unaided decisions in similar cases. Further research is needed to establish if the results of our study can be applied to a broader range of ecological tasks and different contexts. Despite this limitation we believe our findings are significant due to their novelty and potential to inspire future research across multiple disciplines.

5. Conclusion

Our study explored the potential impact of machines inadvertently acting as mentors on the knowledge acquisition and skill development of novices and students. We focused specifically on three key aspects: bias transmission, bias generalization, and the upskilling effect. We believe that our work is particularly important for understanding the risks associated with bias transmission and bias generalization in interactions with machines. In our experiment, we observed how a fictitious artificial intelligence transmitted its systematic error, or bias, to students. This bias persisted even after the AI support was removed and generalized to new ambiguous cases.

The results of our user study highlight the potential for clinical decision support systems (DSS) to act as "bad mentors", inadvertently conveying incorrect decision criteria to users who may not be aware of the machines' errors. In an unaided setting following a relatively short interaction with AI, the substantial effects observed in the group that interacted with AI compared to the control group suggest the need for further studies to investigate this phenomenon in greater depth, as well as more in general the effect of the introduction of accurate DSS in the learners' practice.

Future research should include diverse samples of participants, engage them in learning conditions that reflect more realistic criteria (e.g., integrating multiple and varied visual cues), and assess their competence after an appropriate washout period to evaluate the long-term persistence of the effects.

CRediT authorship contribution statement

Lucia Vicente: Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Helena Matute:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Caterina Fregosi:** Writing – review & editing, Writing – original draft, Formal analysis. **Federico Cabitza:** Writing – review & editing, Writing – original draft, Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT Plus (GPT-4o) by OpenAI in order to enhance the English fluency. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

LV, FC and CF acknowledge funding support provided by the Italian project PRIN PNRR 2022 InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision-making. CUP: H53D23008090001 funded by the European Union - Next Generation EU. HM acknowledges funding support from Grant PID2021-126320NB-I00 funded by MCIN/AEI/10.13039/501100011033 and ERDF A Way of Making Europe.

Data availability

The data that support the findings of this study are freely available on the Open Science Framework at https://osf.io/uxy79/.

References

Adam, H., Balagopalan, A., Alsentzer, E., Christia, F., Ghassemi, M., 2022. Mitigating the impact of biased artificial intelligence in emergency decision-making. Commun. Med. 2, 149.

Agudo, U., Liberal, K.G., Arrese, M., Matute, H., 2024. The impact of AI errors in a human-in-the-loop process. Cogn. Res.: Princ. Implic. 9, 1.

Bandura, A., Walters, R.H., 1977. Social Learning Theory, vol. 1, Prentice Hall, Englewood Cliffs, NJ.

Beane, M., 2019. Learning to work with intelligent machines. Harv. Bus. Rev. 97 (5), 140–148.

Blanco, F., Moreno-Fernández, M.M., Matute, H., 2020. Are the symptoms really remitting? How the subjective interpretation of outcomes can produce an illusion of causality. Judgm. Decis. Mak. 15 (4), 572–585.

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T.F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T.L., Henrich, J., et al., 2023. Machine culture. Nat. Hum. Behav. 7 (11), 1855–1868.

Cabitza, F., Campagner, A., Angius, R., Natali, C., Reverberi, C., 2023a. AI shall have no dominion: on how to measure technology dominance in AI-supported human decision-making. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–20.

Cabitza, F., Campagner, A., Ronzio, L., Cameli, M., Mandoli, G.E., Pastore, M.C., Sconfienza, L.M., Folgado, D., Barandas, M., Gamboa, H., 2023b. Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. Artif. Intell. Med. 138, 102506.

Cabitza, F., Locoro, A., Alderighi, C., Rasoini, R., Compagnone, D., Berjano, P., 2019. The elephant in the record: on the multiplicity of data recording work. Heal. Inform. J. 25 (3), 475–490.

Cabitza, F., Rasoini, R., Gensini, G.F., 2017. Unintended consequences of machine learning in medicine. Jama 318 (6), 517–518.

Castaneda, C., Nalley, K., Mannion, C., Bhattacharyya, P., Blake, P., Pecora, A., Goy, A., Suh, K.S., 2015. Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. J. Clin. Bioinform. 5, 1–16.

Chen, J.H., 2024. Who's training whom? NEJM AI 1 (5).
Dratsch, T., Chen, X., Mehrizi, M.R., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., dos Santos, D.P., 2023. Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology 307, e222176.
Fügener, A., Grahl, J., Gupta, A., Ketter, W., 2022. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. Inf. Syst. Res. 33 (2), 678–696.
Gaessler, F., Piezunka, H., 2023. Training with AI: Evidence from chess computers. Strat. Manag. J. 44 (11), 2724–2750.
Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S.J., Lermer, E., Coughlin, J.F., Guttag, J.V., Colak, E., Ghassemi, M., 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Digit. Med. 4 (1), 31.
Gero, K.I., Ashktorab, Z., Dugan, C., Pan, Q., Johnson, J., Geyer, W., Ruiz, M., Miller, S., Millen, D.R., Campbell, M., Kumaravel, S., Zhang, W., 2020. Mental models of AI agents in a cooperative game setting. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–12.
Glickman, M., Sharot, T., 2023. How human-AI feedback loops alter human perceptual, emotional and social judgements. OSF Prepr..
Glickman, M., Sharot, T., 2024. AI-induced hyper-learning in humans. Curr. Opin. Psychol. 60, 101900.
Goddard, K., Roudsari, A., Wyatt, J.C., 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J. Am. Med. Inform. Assoc. 19 (1), 121–127.
Green, B., Chen, Y., 2019. The principles and limits of algorithm-in-the-loop decision making. Proc. ACM Hum.-Comput. Interact. 3 (CSCW).
Henrich, J., 2016. The Secret of Our Success: How Culture is Driving Human Evolution, Domesticating Our Species, and Making us Smarter. Princeton University Press.
Henrich, J., Gil-White, F.J., 2001. The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. Evol. Hum. Behav. 22 (3), 165–196.
Heyes, C., 2017. When does social learning become cultural learning? Dev. Sci. 20 (2), e12350.
Heyes, C., 2018. Cognitive Gadgets: The Cultural Evolution of Thinking. Harvard University Press.
Holzinger, A., Saranti, A., Angerschmid, A., Finzel, B., Schmid, U., Mueller, H., 2023. Toward human-level concept learning: Pattern benchmarking for AI algorithms. Patterns 4 (8).
Jacobs, M., Pradier, M.F., McCoy, T.H., Perlis, R.H., Doshi-Velez, F., Gajos, K.Z., 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Transl. Psychiatry 11 (1), 108.
Kahneman, D., 2011. Fast and Slow Thinking. Allen Lane Penguin Books, NewYork.
Kupfer, C., Prassl, R., Fleiß, J., Malin, C., Thalmann, S., Kubicek, B., 2023. Check the box! How to deal with automation bias in AI-based personnel selection. Front. Psychol. 14, 1118723.
Kwong, J.C.C., dan Nguyen, D., Khondker, A., Kim, J.K., Johnson, A.E.W., McCradden, M.M., Kulkarni, G.S., Lorenzo, A., Erdman, L., Rickard, M., 2024. When the model trains you: Induced belief revision and its implications on artificial intelligence research and patient care — A case study on predicting obstructive hydronephrosis in children. NEJM AI 1, 1–7.
Lai, V., Chen, C., Smith-Renner, A., Liao, Q.V., Tan, C., 2023. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. FAccT'23, Association for Computing Machinery, New York, NY, USA, pp. 1369–1385.
Nass, C., Moon, Y., 2000. Machines and mindlessness: Social responses to computers. J. Soc. Issues 56 (1), 81–103.
Nonaka, I., Takeuchi, H., 2007. The knowledge-creating company. Harv. Bus. Rev. 85 (7/8), 162.
Parasuraman, R., Manzey, D.H., 2010. Complacency and bias in human use of automation: An attentional integration. Hum. Factors 52 (3), 381–410.
Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J., 2022. AI in health and medicine. Nature Med. 28 (1), 31–38.
Reason, J., 1990. Human Error. Cambridge University Press.
Rezazade Mehrizi, M.H., Mol, F., Peter, M., Ranschaert, E., Dos Santos, D.P., Shahidi, R., Fatehi, M., Dratsch, T., 2023. The impact of AI suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. Sci. Rep. 13, 9230.
Shin, M., Kim, J., van Opheusden, B., Griffiths, T.L., 2023. Superhuman artificial intelligence can improve human decision-making by increasing novelty. Proc. Natl. Acad. Sci. 120, 1–6.
Skinner, B.F., 1958. Teaching machines: From the experimental study of learning come devices which arrange optimal conditions for self-instruction. Science 128 (3330), 969–977.
Skinner, B.F., 1961. Why we need teaching machines. Harv. Educ. Rev. 31, 377–398.
Tejeda, H., Kumar, A., Smyth, P., Steyvers, M., 2022. AI-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies. Comput. Brain Behav. 5 (4), 491–508.
Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nature Med. 25 (1), 44–56.
Tozsin, A., Ucmak, H., Soyturk, S., Aydin, A., Gozen, A.S., Fahim, M.A., Güven, S., Ahmed, K., 2024. The role of artificial intelligence in medical education: A systematic review. Surg. Innov. 31 (4), 415–423.
Tractinsky, N., Katz, A.S., Ikar, D., 2000. What is beautiful is usable. Interact. Comput. 13 (2), 127–145.
Vaccaro, M., Almaatouq, A., Malone, T., 2024. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nat. Hum. Behav. 8 (12), 2293–2303.
Vered, M., Livni, T., Howe, P.D.L., Miller, T., Sonenberg, L., 2023. The effects of explanations on automation bias. Artificial Intelligence 322, 103952.
Vicente, L., Matute, H., 2023. Humans inherit artificial intelligence biases. Sci. Rep. 13 (1), 15737.
Yu, K., Berkovsky, S., Taib, R., Zhou, J., Chen, F., 2019. Do I trust my machine teammate? an investigation from perception to decision. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. IUI '19, Association for Computing Machinery, New York, NY, USA, pp. 460–468.