Collective and Individual Decisions
Vol. 37, No. 1, 2025, pp. 57–80, e-ISSN: 3071-7973

No Algorithm Aversion in Improving AI

Wojciech Milczarski[1], Anna Borkowska[2], Emilia Biesiada[3], Laura Russak[4], Michał Białek[5]

Abstract

Algorithm aversion is the tendency to avoid using algorithms or AI systems. Most of the studies only present participants with the most recent performance of humans or AI. Through a series of experiments, involving N=905 participants, we investigated how people evaluate AI versus human performance, focusing on consistent high-quality output versus improvement over time. Our findings reveal that the preference for humans over AI significantly decreases when both demonstrate improvement. We observe this in the creative domain of tattoo design and other non-creative domains, such as law, logistics, and sales. Emphasizing AI's improvement could effectively reduce algorithm aversion and increase AI acceptance. Our research contributes to understanding AI acceptance in domains, even those traditionally dominated by human creativity, offering insights for implementing AI systems in various fields.

Keywords: Artificial intelligence, creativity, tattoo design, performance evaluation, algorithm aversion.

[1] Wojciech Milczarski – corresponding author, Institute of English Studies, The Faculty of Languages, Literatures, and Cultures, University of Wrocław, Wrocław, Poland, e-mail: wojciech.milczarski@uwr.edu.pl; ORCID: https://orcid.org/0000-0002-8215-5190

[2] Anna Borkowska – Institute of Psychology, The Faculty of Historical and Pedagogical Sciences, University of Wrocław, Poland, ORCID: https://orcid.org/0000-0001-5428-0974

[3] Emilia Biesiada – Institute of Psychology, The Faculty of Historical and Pedagogical Sciences, University of Wrocław, Poland, ORCID: https://orcid.org/0000-0002-0113-7290

[4] Laura Russak – Institute of Psychology, The Faculty of Historical and Pedagogical Sciences, University of Wrocław, Poland, ORCID: https://orcid.org/0000-0002-8177-5857

[5] Michał Białek – Institute of Psychology, The Faculty of Historical and Pedagogical Sciences, University of Wrocław, Poland, ORCID: https://orcid.org/0000-0002-5062-5733

Introduction

Imagine walking into a tattoo shop, excited to get a design you have dreamed about for years. The artist pulls out a tablet, taps the screen a few times, and shows you a stunning, personalized tattoo created by an AI in seconds. How would you feel? Amazed at the technology? Unsettled by the lack of human touch? Or perhaps a mix of both? This scenario is no longer science fiction. AI-generated art is quickly advancing with systems like DALL-E and Midjourney creating images that rival human artists. As AI develops in creative domains once thought to be uniquely human, it raises questions about our relationship with technology.

Tattoo design is an interesting case for studying these questions. Tattoos are permanent, deeply personal, and have emotional and cultural significance. The idea of an AI designing something this intimate and lasting challenges what we think about art and human creativity.

In this article, we discuss the overlap of AI, creativity, and performance perception. We investigate how people evaluate AI performance in tattoo design alongside other fields compared to human agents, exploring whether we may observe previously established human judgment biases in AI.

Previous research has shown that people often avoid using algorithms for decision-making, even after seeing them perform better than humans (Dietvorst *et al*., 2015). This is known as "algorithm aversion." Most of the studies on this phenomenon have focused on objective domains like hiring decisions (Kuncel, 2008) or medical diagnosis (Grove *et al*., 2000). Human attitudes toward AI are less understood in creative domains. In their recent work on AI creativity, Magni *et al.* (2024) found that people sometimes, but not always, perceive AI-created products as less creative than human-created ones. This bias was stronger for artistic works (e.g., paintings) and weaker for business-related items (e.g., advertisements or start-up ideas). Importantly, perceptions of effort played a significant role in this bias, with people consistently attributing less effort to AI-created work. This suggests that there might be something special about how we perceive creativity and effort in AI systems. Our study builds on this research, testing whether similar biases exist in tattoo design.

AI aversion research focuses almost entirely on recent performance. However, when evaluating effort in performance, people often consider both initial and recent performance. Studies have investigated how people judge two different types of performance trajectory: consistently good performance versus performance that starts lower but improves over time to reach the same level (Soliman & Buehler, 2018a, 2018b). This research has shown that people perceive greater effort in candidates who show improvement, viewing them as more deserving. This tendency persisted even when the overall performance of consistent candidates was objectively superior (Soliman & Buehler, 2018a, 2018b).

The relative weight given to future expectations and performance evaluations varies depending on the decision context. In contexts focused on future performance (e.g., job promotions or hiring decisions), respondents favored improving candidates. In contrast, in achievement-focused contexts (e.g., academic scholarships or rewards), respondents preferred consistently performing candidates. These findings highlight the complexity of performance evaluation and the importance of considering both the performance trajectory and the specific evaluation context.

 DOI: 10.7206/cid.3071-7973.9

This finding aligns with previous research on the importance of perceived effort in performance evaluations (Soliman & Buehler, 2018a; 2018b) and raises interesting questions about how attribution of effort might differ between human and AI agents. Given these complex and sometimes contradictory findings, we aimed to explore how these different factors interact in the context of AI-produced creative work, specifically in the context of tattoo artistry. We will examine how people evaluate tattoos created by AI systems with different performance trajectories: those that show consistent high-quality output versus those that demonstrate improvement over time. We operationalized improvement as performance that started below and ended at the exact level of the corresponding constant score. For example, we put together a constant "7" out of "10" performance, together with a progress performance going from "4" to "7" out of "10."

By focusing on tattoos, a form of art that is both personal and permanent, we aimed to investigate whether observing improvement, which influences judgments in other domains, has the same effect for AI-created art. Furthermore, we wanted to identify any potential differences in how evaluators approach AI performance compared to human performance in artistic contexts.

Our study consists of a series of experiments designed to address these key questions: i) How do people evaluate consistent high-quality performance versus improvement over time in tattoo design? ii) Are there differences in these evaluations between human artists and AI systems? By addressing these questions, we hoped to gain informative insights into the development and implementation of AI systems in creative domains.

Experiments overview

Across a Pilot study and four main experiments with a total N=905, we investigated how people evaluate the performance of AI and humans. The first study was exploratory, whereas follow-up studies (2–4) were preregistered, and served the goal of establishing the robustness of the findings and testing their alternative explanations. The logic was sound, as exploratory research is important for the pursuit of knowledge (Aczel, 2024).

In all studies, our participants were native English speakers from the United Kingdom recruited via the Prolific online platform. We conducted all analyses using Jamovi version 2.2 (The jamovi project, 2021), with the use of the GAMLj package (Gallucci, 2019).

Data and materials for all of the experiments are available at:
https://osf.io/uvdpx/?view_only=bd8f21a68f924857bb8da28bd444d99b
Experiment 2 was pre-registered: https://aspredicted.org/5G5_VTN.
Experiment 3 was pre-registered: https://aspredicted.org/Z67_TRN.
Experiment 4 was pre-registered: https://aspredicted.org/G84_XYR.

Pilot experiment

We conducted a Pilot study to select appropriate materials for our main experiments. This step was important to ensure that the tattoo images we used in the following studies accurately represented varying quality levels.

Participants

In the pilot experiment, we recruited N=59 participants (mean age=36.8, age range= 22–67, 26 women, 33 men) through the Prolific platform. To get relevant background information, we asked participants about their personal tattoo experience. In our sample, N=42 reported having no tattoos, N=8 had exactly one tattoo, and N=9 had more than one tattoo.

Procedure and Materials

We presented participants with 150 pictures of tattoos of different quality and style on various body parts. We selected these tattoos from online sources (e.g., Google Images, Pinterest) to represent a range of quality levels while reducing the influence of tattoo style on liking ratings. To achieve this, we focused on black ink tattoos with a similar art and line visual style.

Participants rated each tattoo on a scale from 1–10 using a slider without gridlines. To reduce the task length, each participant saw 30 randomly selected photos out of the total database of 150 tattoos. This allowed us to gather ratings on all tattoos while keeping the task manageable for individual participants. On average, participants spent six minutes completing the study.

Results

We used the data collected from the pilot study to calculate a mean liking score for each of the 150 tattoos. Afterwards, two independent experts reviewed the tattoos, considering both the numeric scores (means and standard deviations of the ratings) and qualitative factors such as the styles of the tattoos and the overall photo quality. Based on this evaluation, we selected twelve tattoos for the main experiments: nine representing good quality work and three representing lower quality work. Readers may find chosen photos on the OSF page of the project.

Experiment 1

We designed this experiment to directly investigate how people evaluate the performance of human artists and AI systems in the context of tattoo design. Specifically, we compared ratings of consistent high-quality performance versus improvement over time. This experiment addressed key questions about performance evaluations in creative fields and potential differences in perceptions of human versus AI-generated art.

Participants

For Experiment 1, we recruited N=208 participants through the Prolific platform (Mean Age=41.8, Age Range=21–76, 97 female, 108 male, 3 others). We excluded additional N=4 participants who failed an attention check. To get relevant background information, we asked participants about their personal experiences with tattoos. In the sample, N=108 reported having no tattoos, N=40 had exactly one tattoo, and N=60 had more than one tattoo. There were also N=5 professional tattoo artists among the sample.

The attention check used in this experiment asked participants, “How many times did you have a fatal heart attack while watching Netflix?” We excluded from the study all the participants who gave any answer other than “0.”

 DOI: 10.7206/cid.3071-7973.9

Procedure and materials

We presented participants with pairs of tattoo images, each pair representing either consistent high-quality work or improvement over time. We selected the images based on the ratings obtained in the pilot study, ensuring that they accurately represented varying levels of tattoo quality.

We presented each pair of tattoos horizontally in a separate question. The left image represented the artist's or AI's earlier work, while the right image showed their most recent work (see Figure 1). We chose this layout to clearly communicate the performance trajectory over time, as humans perceive items presented on the left as older and items presented on the right as newer.[6] For the constant performance condition, both images were piloted to ensure participants perceived them as of similar high quality. For improving performance conditions, the earlier work was of lower quality, while the recent work matched the quality of the constant performance condition. The exact wording of an exemplary task was as follows:

> **Human condition:** Your friend is an owner of a tattoo studio. They want to hire a new tattoo artist and show you the resume of the candidate, containing the picture of his work from the beginning of his career (column "the beginning of work") and the picture of his most recent work (column "most recent work"). Please choose, by dragging a slider through the range, if you would recommend that artist to your friend.
>
> **AI condition:** Your friend is an owner of a tattoo studio. They want to buy an AI machine making tattoos, and they show you the first (column "the beginning of work") and the most recent (column "most recent work") work of the machine they consider buying. Please choose, by dragging a slider through the range, if you would recommend that machine to your friend.
>
> In each of the questions concerning AI, we defined the term:
>
> Artificial Intelligence (AI) – the ability of machines to demonstrate skills such as reasoning, learning, planning and creativity. Artificial intelligence enables technical systems to perceive their environment, deal with what they perceive and solve problems, and learn by acting towards a specific goal and improving performance.

Respondents provided ratings using a slider that ranged from "definitely not recommend" to "definitely recommend" without numerical values displayed. We did not use numerical values to prevent participants from trying to assign a specific numeric score and encourage them to focus on their overall impression. We randomly assigned each participant to evaluate either human artists or AI systems.

[6] We presented stimuli in fixed chronological order to avoid potential confusion in more careless participants answers. Mental line (left – earlier, right – later) is generally consistent with the idea that we are culturally wired for the order of our writing system while representing the chronological flow of time or the mental number scales (see among others Patro & Haman, 2012; Borkowska & Milczarski, 2021). Violating this order might result in weakening the potential true effects (Sawicki & Białek, 2017).

FIGURE 1. Layout of the Experimental Tasks in Experiment 1

Source: Tattoo on the left: https://fr.pinterest.com/pin/tattoo-par-willem-tattoo-engraved-tatouage-blackartist-engraving--blacktattoing-black-blackwork-blacktattooart-cannes-sangpiternel-noir--143130094388185098; Tattoo on the right: https://www.worldtattoogallery.com/post/100012985/owl-with-skull-tattoo-by-steve-butcher

Results

We conducted a mixed model analysis. The dependent variable was the recommendation judgment, with the agent (human vs. AI) as a between-subject factor and trajectory of performance (constant performance vs. progress) as a within-subject factor. Both of the factors were coded as -0.5 (AI and constant) and 0.5 (human and progress). We used participant ID and item number as clustering variables to account for individual differences and item-specific effects.

The analysis showed a significant main effect of trajectory of performance (p=.033) with participants preferring constant performance over progress. The main effect of Agent was not significant (p=.470), suggesting that participants did not have an overall preference for human artists over AI systems. The interaction between those factors was also not significant (p=.146). See Table 1 and Figure 2 for detailed results and their visualization.

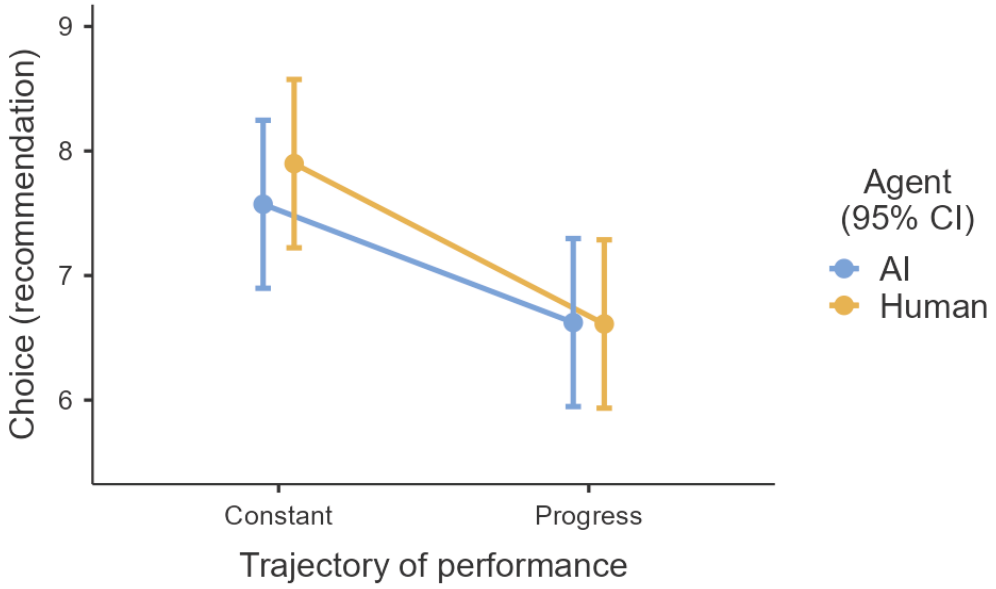
TABLE 1. Results of Experiment 1 in the Mixed Model

	Recommendation (1–10)
$N_{observations}$	1248
Agent (human – AI)	0.16 [-0.27, 0.58]
Trajectory of performance (progress – constant)	**-1.12 [-1.80, -0.44]***
Agent x trajectory of performance	-0.34 [-0.79, 0.12]
Participant ID	ICC=0.29
Item number	ICC=0.04
R2 conditional	0.35
R2 marginal	0.05

Note. *p<.05, **p<.01, ***p<.001.

Source: own elaboration.

FIGURE 2. Plot of the Interaction of the Agent and Trajectory of Performance in Experiment 1



Source: own elaboration.

Discussion

In Experiment 1, we found a preference for constant performance over progress in tattoo design. This was incongruent with previous literature on performance evaluation, which generally suggests that people prefer progress in performance when assessing candidates for jobs (Soliman & Buehler, 2018a; 2018b). This unexpected finding suggests that the evaluation of creative work, or at least of tattoo design, might differ from evaluations in other fields. Surprisingly, we found no effect of the agent (i.e., AI did not significantly differ from humans), showing no algorithm aversion. These findings are not in line with previous algorithm aversion studies in creative domains, suggesting that when judging tattoo designs, people might focus more on the end product rather than its creator.

To validate our findings, we designed Experiment 2 to see whether the preference for constant versus improving performance would differ across different domains.

Experiment 2

The unexpected results of Experiment 1, which showed a preference for constant performance over improvement in a hiring context, led us to design Experiment 2. This experiment was supposed to address the incongruence with previous literature by comparing preference for constant versus improving performance across different domains, including sales, law, and logistics, alongside tattoo design. We wanted to investigate if the preference for constant performance was specific to tattoo design or if it would generalize to other fields.

In this experiment, we made a significant change to our methodology by representing performance with graphs rather than photos. This change was necessary to allow for comparison across different domains and to match our methodology more closely with previous research on the issue.

Participants

For Experiment 2, we recruited N=195 (mean age=40.6, age range=18–73, 125 female, 70 male) through the Prolific platform. We excluded additional N=5 participants based on failed attention checks. The attention checks used in this experiment were more comprehensive than in Experiment 1. We asked participants both "How many times did you have a fatal heart attack while watching Netflix?" and "What is your favorite food? (Please choose the answer "Pancakes".)" We excluded from the analysis participants who gave any answer other than "0" to the first question or any answer other than "Pancakes" to the second question.

Procedure and materials

Participants saw a series of questions, each accompanied by a graph. Similarly to Experiment 1, the context was that a company was looking for a candidate to fill in a position. The graph always presented the performance of one supposed candidate over time (see Figure 3 for an example). We always operationalized constant performance as a line on a given level, whereas progressing performance started at a lower point of the scale, but always ultimately reached the constant performance level (allowing the participants to think that, at the time of the judgment, performance in both scenarios is the same).

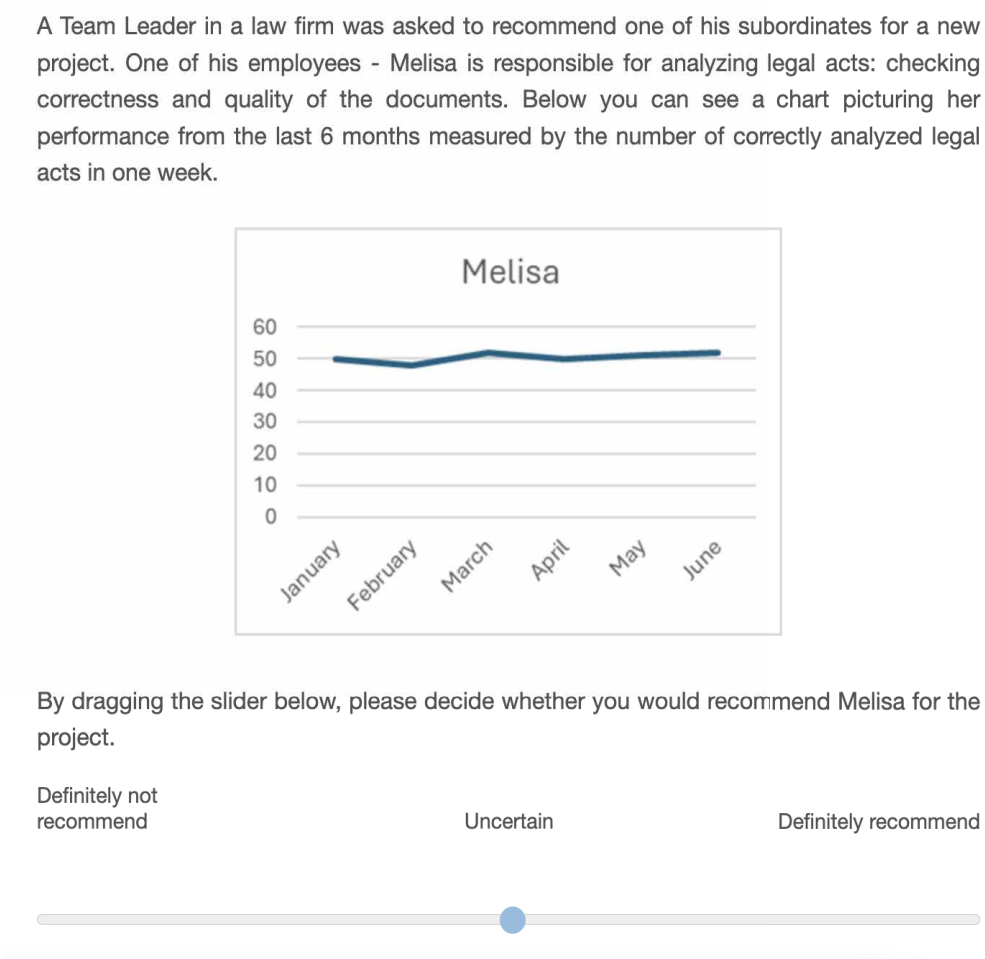
The design was 2 (agent: human, AI - between-subject) x 2 (trajectory of performance: constant, progress – within-subject). Half of the participants evaluated human candidates, while the other half evaluated AI models. For each participant, half of the graphs showed consistent performance, while the other half showed improvement over time.

We included four types of careers in the stimuli: sales, logistics, law, and tattoo design. Including tattoo design allowed us to obtain a more objective measure of preference for constant/improving performance in the same domain we explored in our previous experiment.

Participants' task was to decide how likely they would be to recommend a candidate for a job/task. They responded by dragging a slider with no numerical values displayed, ranging from "definitely not recommend" to "definitely recommend," with "uncertain" in the middle. This allowed for specific judgments without forcing participants to use numeric values.

 DOI: 10.7206/cid.3071-7973.9

FIGURE 3. Example Stimulus from Experiment 2



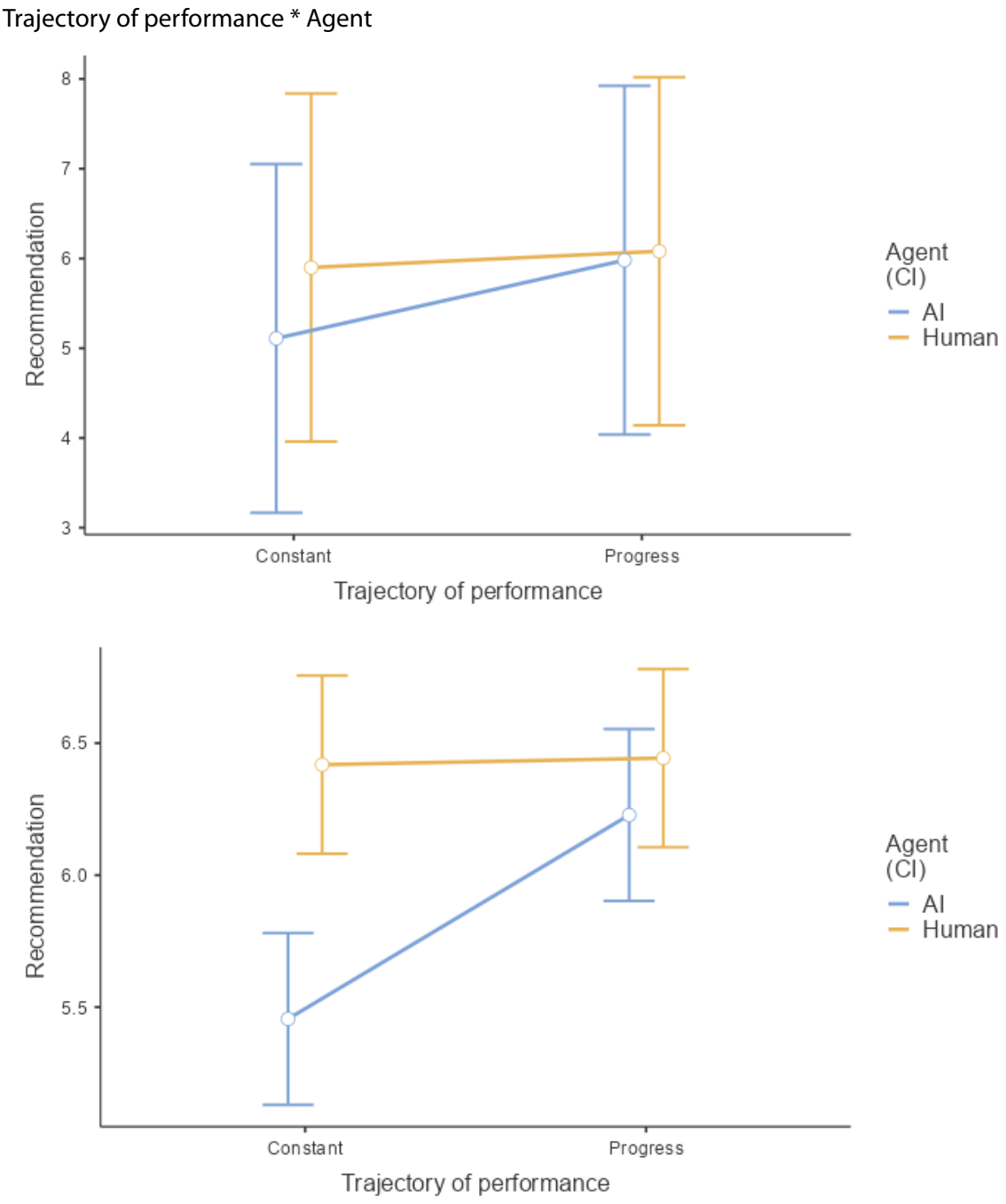
Note. For the AI condition, we used unique names for each model, e.g., *Blizram2000*.

Source: own elaboration.

Results

We conducted a mixed model analysis with the recommendation judgment as the dependent variable. Agent (human vs. AI) was a between-subject factor, and trajectory of performance (constant performance vs. progress) was a within-subject factor. We coded both of the factors as -0.5 (AI and constant) and 0.5 (human and progress). We used participant ID and type of job (logistics/law/sales/tattoo design) as clustering variables (Table 2, M1).

FIGURE 4. Plot of the Interaction Agent x Trajectory of Performance: Experiment 2 Results in M1 (Panel A), and M2 (Panel B)



Source: own elaboration.

Unlike in Experiment 1, we found a main effect of the agent, showing a preference toward humans over AI. However, this effect was qualified by an interaction: this preference persisted when the agent showed constant performance, but disappeared when the agent improved (see Figure 4).

To see whether tattoos were perceived differently from other fields, we conducted an additional, exploratory analysis that included a new variable: tattoos (tattoos vs. non-tattoos, see M2, Table 2) rather than having job type as a random effect. This factor was coded -0.5

 DOI: 10.7206/cid.3071-7973.9

(non-tattoos) and 0.5 (tattoos). This analysis showed three significant interactions. The first was agent by tattoos: people showed a preference for humans in tattoo design, while there was little difference between humans and AI in other domains. The second observed interaction was trajectory of performance by tattoos: while there was a significant preference for progress in non-tattoo domains, there was no preference in the tattoo domain. The third observed interaction was agent by trajectory of performance. When describing a constant performance, participants preferred humans over AI. We did not observe any such preferences in the case of progressing performance.

TABLE 2. Results of Experiment 2 in the Mixed Model

	Recommendation (1–10)	
	M1 (confirmatory)	M2 (exploratory)
$N_{observations}$	1560	1560
Agent (human – AI)	**0.45 [0.07, 0.82]***	**0.59 [0.19, 0.99]****
Trajectory of performance (progress – constant)	**0.53 [0.33, 0.72]*****	**0.40 [0.15, 0.65]****
Agent x trajectory of performance	**-0.69 [-1.08, -0.30]*****	**-0.75 [-1.24, -0.25]****
Tattoos (tattoos – non-tattoos)		**1.48 [1.23, 1.72]*****
Agent x tattoos		**0.58 [0.08, 1.07]***
Trajectory of performance x tattoos		**-0.51 [-1.01, -0.02]***
Agent x trajectory of performance x tattoos		-0.23 [-1.22, 0.76]
Participant ID	ICC=0.26	ICC=0.21
Job type	ICC=0.29	-
R2 conditional	0.45	0.28
R2 marginal	0.02	0.09

Note. *p<.05, **p<.01, ***p<.001.

Source: own elaboration.

We also plotted the agent by trajectory of performance interaction (Figure 4, Panel B). The results followed the pattern we previously observed; participants preferred humans over AI in the constant condition, but they did not have a preference in the progress condition.

Discussion

The results of Experiment 2 differed from our findings in Experiment 1, aligning more closely with previous literature that emphasized the value of progress in performance. This change likely happened for two reasons. Firstly, by introducing domains other than tattoo design, we could see how preferences for constant vs. improving performance varied across different professional contexts. Secondly, the use of performance graphs rather than photos may have influenced how participants perceived and judged performance trajectories.

The preference for humans in tattoo design contrasted with little difference between humans and AI in other domains, which would align with a stronger bias against AI in artistic contexts (Magni *et al*., 2024). It is possible that people believed tattoo design required uniquely human traits that they did not associate with AI.

The agent by trajectory of performance interaction is particularly interesting from an AI aversion perspective. When judging candidates with constant performance, people showed AI aversion. However, when judging candidates with improvement over time, this aversion disappeared. The finding suggests a potential strategy for mitigating AI aversion. Showcasing an AI's ability to improve might be an effective way of increasing its acceptance and perceived value, especially in domains where initial AI aversion is strong.

The lack of preference for progress in the tattoo domain, contrasted with the presence of the preference for progress in other domains, reinforces the idea that tattoo design might be evaluated differently from other fields. This could be due to the permanent nature of tattoos, leading to a higher value placed on consistency and reliability.

The shift in results from Experiment 1 to Experiment 2 suggests that the way of presenting performance (photos vs. graphs) may influence performance perceptions.

These findings led us to believe that people might see certain domains differently in terms of the role of innate talent versus learned skill. We speculated that the difference in preferences between tattoo design and other domains might be driven by people's beliefs about whether performance in each is mostly predicted by talent or by learning and practice. This hypothesis formed the basis for our next experiment, which aimed to further unpack AI aversion in various domains.

Experiment 3

The results of Experiment 2 led us to speculate that people perceive certain domains differently in terms of the importance of innate talent versus learned skill. To directly test this hypothesis and its potential relationship to AI aversion, we designed Experiment 3 to explicitly measure how participants attribute performance to innate talent vs. learned skill across the previously used domains. This experiment was necessary to help us understand the reasons for the differences we saw between tattoo design and other fields.

Participants

For Experiment 3, we recruited N=52 participants (Mean Age=33.9, Age Range=18–68, 36 women, 15 men, 1 other) through the Prolific platform. We excluded additional N=3 participants who failed attention checks. The attention checks used in this experiment were the same as in Experiment 2.

Procedure and Materials

We presented participants with a series of questions, each accompanied by a graph showing performance over time. The instructions stated that all of the graphs showed good performance of human agents; half of the graphs showed consistent performance, and the other half showed improvement over time.

We used a within-subject design, with participants seeing scenarios from two conditions: creative domain (tattoo design) and other domains (sales, logistics, law), same as in Experiment 2. For each scenario, we asked participants to assess the degree to which innate talent or learned skills were responsible for the performance presented.

Participants gave their judgements using a single Likert scale, marking their answers by dragging a slider from 1 (innate talent) to 10 (learned skill). This allowed us to get specific

 DOI: 10.7206/cid.3071-7973.9

judgments about the perceived source of performance in each domain and for each performance trajectory.

Results

We conducted a mixed model analysis with the judgment on the source of performance (1 – innate talent, 10 – learned skill) as the dependent variable. The random effect was the domain (non-creative vs. creative). It was coded -0.5 (creative) and 0.5 (non-creative). Participant ID and trajectory of performance were clustering variables (Table 3, M1).

TABLE 3. Results of Experiment 3 in the Mixed Model

	Source of performance (1 – innate talent, 10 – learned skill)	
	M1 (confirmatory)	M2 (exploratory)
$N_{observations}$	208	208
Domain (non-creative – creative)	-0.18 [-0.85, 0.49]	-0.19 [-0.86, 0.49]
Trajectory of performance (progress – constant)		**2.28 [1.59, 2.98]*** **
Domain x trajectory		0.41 [-1.0, 1.81]
Participant ID	ICC=0.10	ICC=0.09
Trajectory of performance (progress – constant)	ICC=0.38	
R2 conditional	0.42	0.30
R2 marginal	<.001	0.22

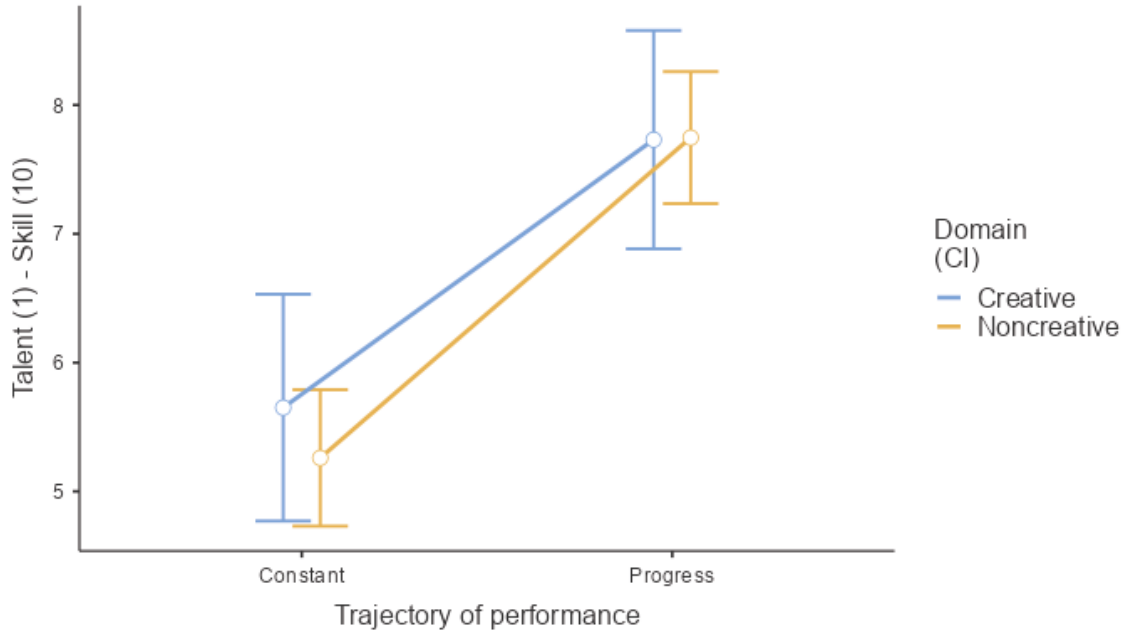
Note. *p<.05, **p<.01, ***p<.001.

Source: own elaboration.

We found no inherent differences in how participants perceived the performance source (innate talent vs. learned skill) between creative (tattoo design) and non-creative (law/logistics/sales) domains (M1, Table 3).

However, when we explored whether there was a main effect of progress vs. constant performance on the perceived source of that performance, we found a significant effect (M2, Table 5 and Figure 5). We coded the effect as -0.5 (Constant) and 0.5 (Progress). Participants perceived progress in performance as predominantly stemming from learned skill (mean=7.74), whereas for constant performance, they perceived the source of performance as falling somewhere in the middle between talent and skill (mean=5.45). The interaction between Domain and Trajectory was not significant.

FIGURE 5. Experiment 3 Results for M2



Source: own elaboration.

Discussion

Contrary to our expectations, we found no significant differences in how participants attributed performance to innate talent versus learned skill between creative (tattoo design) and non-creative domains. This suggests that we cannot explain the differences we observed in preferences for constant vs. improving performance across domains in Experiment 2 by different perceptions of the role of talent vs. skill in these fields. From an AI aversion perspective, this finding is particularly interesting as it suggests that people may not inherently view creative domains as more reliant on innate, "human" qualities that AI might lack.

Instead, we found a robust effect of performance trajectory on perceptions of the source of performance. This effect provides important insights into how people interpret improvement in both human and AI performance. Regardless of the domain, when people saw performance improvement, they were more likely to believe it stemmed from learned skills. Conversely, consistent performance was more likely to be seen as a mix of innate talent and learned skills.

The association of improvement with learned skills suggests that people are open to the idea of AI systems developing and improving their capabilities over time, rather than seeing them as static and pre-programmed. This could potentially reduce AI aversion, perhaps by highlighting qualities (effort, learning) that are typically associated with human performance. Showcasing an AI's ability to learn and improve over time could be effective, especially in domains where initial AI aversion is strong.

Given that we did not find the expected differences in talent vs. skill attribution between domains, we needed to reconsider the reasons for the domain-specific effects observed in Experiment 2. If the perception of performance as stemming from talent or skill is more closely related to the pattern of performance (progress vs. constant) rather than to the specific domain of the performance, why did we see domain-specific effects? We started to wonder whether

 DOI: 10.7206/cid.3071-7973.9

the mode of presentation (pictures of tattoos vs. graphs of performances) might play a more important role in shaping participants' recommendations and perceptions of performance than we had initially thought. This consideration led to the design of our final experiment, which aimed to directly test the role of presentation format in performance evaluations within the creative domain of tattoo design.

Experiment 4

Following our reasoning from the discussion of Experiment 3 and the incongruence between Experiments 1 and 2, we designed Experiment 4 to directly test whether the mode of presentation would influence recommendation scores and preferences for constant vs progressing performance. We speculated that when presented with pictures, participants would prefer constant performance in tattoo design, whereas when presented with graphs, people would prefer progress. This experiment was necessary to help us understand the role of presentation format in shaping judgements of performance trajectories and its potential impact on AI aversion in creative domains.

Participants

For Experiment 4, we recruited N=391 participants (Mean Age=38.5, Age Range=19–84, 245 women, 145 men, 1 other) through the Prolific platform. We excluded additional N=11 participants who failed attention checks. The attention checks were similar to those in Experiments 2 and 3 but with a slight variation. We asked participants about fatal heart attacks while watching Netflix and about their favorite food, but this time, "meat jelly" was the correct answer for the food question.

Procedure and Materials

We presented participants with a series of questions, each accompanied by either a graph or a pair of pictures of tattoos. The context was that a company was looking for a candidate to fill a tattoo designer position. Graphs/pictures always presented the performance of one supposed candidate.

The design was 2x2x2. Half of the participants judged human candidates, while the other half judged AI models. Half of the participants saw graphs, while the other half saw pictures of tattoos. For each participant, half of the tasks showed consistent performance, while the other half showed improvement over time.

We asked participants to decide whether they would recommend the candidate for the job by dragging a slider with no numerical values displayed, ranging from 1 – "definitely not recommend" to 10 – "definitely recommend," with "uncertain" in the middle.

To make sure that the pictures of the tattoos accurately represented the constant/improving performance condition, we included a validation check after each question. Participants had to rate the tattoos, answering the question: "How satisfied do you think the client was with the tattoo on the left/right?" using the slider from 0 ("Not at all") to 10 ("Very much"). For clarity purposes, we provide a table (Table 4) listing all the conditions in the experiment.

TABLE 4. Conditions Used in Experiment 4

	Graph	Pictures
Human	Progress/Constant Human Graph	Progress/Constant Human Pictures
AI	Progress/Constant AI Graphs	Progress/Constant AI Pictures

Source: own elaboration.

Validation

The validation check revealed that participants perceived one of the tattoo picture pairs, which we considered to represent constant performance based on the pilot, as progress (Q_85, Q_70 – they can be found here: https://osf.io/uvdpx/?view_only=035152b3b92f4a1b83ed88b3c26a051d). The mean rating for the "older" tattoo was 6.20, and for the "most recent" tattoo, it was 8.46. This is a difference of 2.26 points, which closely resembles two other point differences in the progress condition (2.54, 2.52). This is why we decided to recode the data for this pair as belonging to the progress condition.

Results

We conducted a mixed model analysis with the recommendation judgment as the dependent variable. The random effects were the agent (human – AI), trajectory of performance (progress – constant), and mode of presentation (pictures – graphs). We coded them as follows: -0.5 (AI; constant; graphs), and 0.5 (human; progress; pictures). We used participant ID and item number as clustering variables (Table 5).

TABLE 5. Results of Experiment 4 in the Mixed Model

	Recommendation (1–10)
$N_{observations}$	2346
Agent (human – AI)	**1.16 [0.85, 1.48]*** **
Trajectory of performance (progress – constant)	-0.16 [-0.94, 0.62]
Mode of presentation (pictures – graphs)	0.31 [-0.76, 1.39]
Agent x trajectory of performance	**-1.18 [-1.49, -0.86]*** **
Agent x mode of presentation	-0.54 [-1.16, 0.09]
Trajectory of performance x mode of presentation	0.70 [-0.86, 2.26]
Agent x trajectory of performance x mode of presentation	**-0.93 [-1.56, -0.30]** **
Participants' ID	ICC=0.36
Item number	ICC=0.20
R2 conditional	0.48
R2 marginal	0.07

Note, *p<.05, **p<.01, ***p<.001.

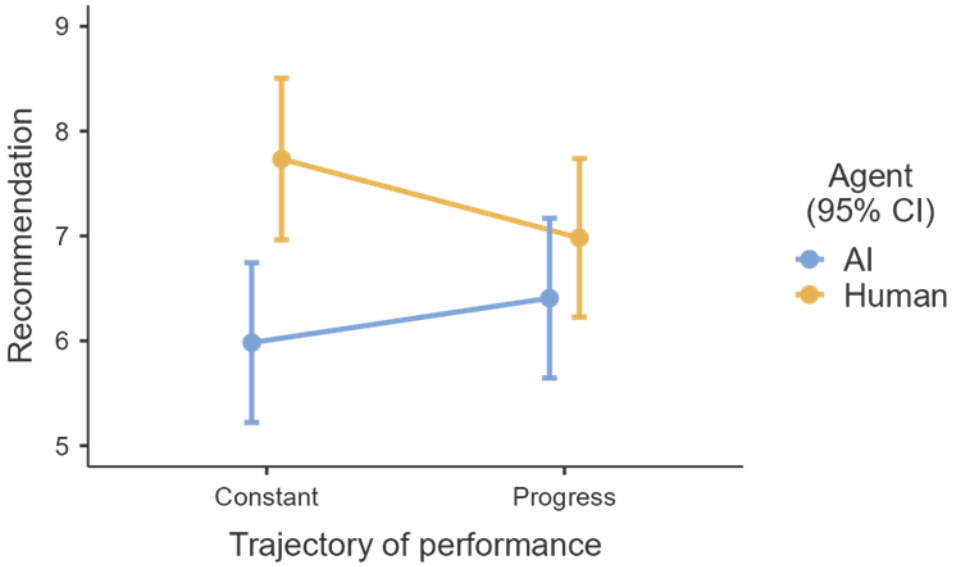
Source: own elaboration.

There was a significant main effect of Agent type, with humans being rated higher overall than AI. We also observed a significant interaction between agent type and trajectory of performance (Figure 6). This interaction suggested that the preference for human agents was

larger in the constant performance condition than in the progress condition. Finally, there was a significant three-way interaction between agent type, trajectory of performance, and mode of presentation ($p<.01$). This complex interaction suggested that the relationships between agent type and performance trajectory differed depending on whether participants saw graphs or pictures.

Contrary to our hypothesis, the mode of presentation (pictures vs. graphs) did not have a significant main effect on recommendations ($p>.5$), and there was no two-way interaction between the mode of presentation and the trajectory of performance. This suggests that the format in which performance was presented did not, on its own, strongly influence participants' evaluations and did not, on its own, cause the discrepancies between Experiments 1 and 2.

FIGURE 6. Plot of the Interaction Agent x Trajectory of Performance: The Results of Experiment 4



Source: own elaboration.

Discussion

The results of Experiment 4 show a set of dependencies between agent type, trajectory of performance, and mode of presentation in shaping judgments of tattoo design performance. Contrary to our initial hypothesis, the mode of presentation (pictures vs. graphs) did not have a significant main effect on recommendations. There was also no two-way interaction between the mode of presentation and the trajectory of performance. This suggests that the format in which performance is presented may not be as important as we had initially thought.

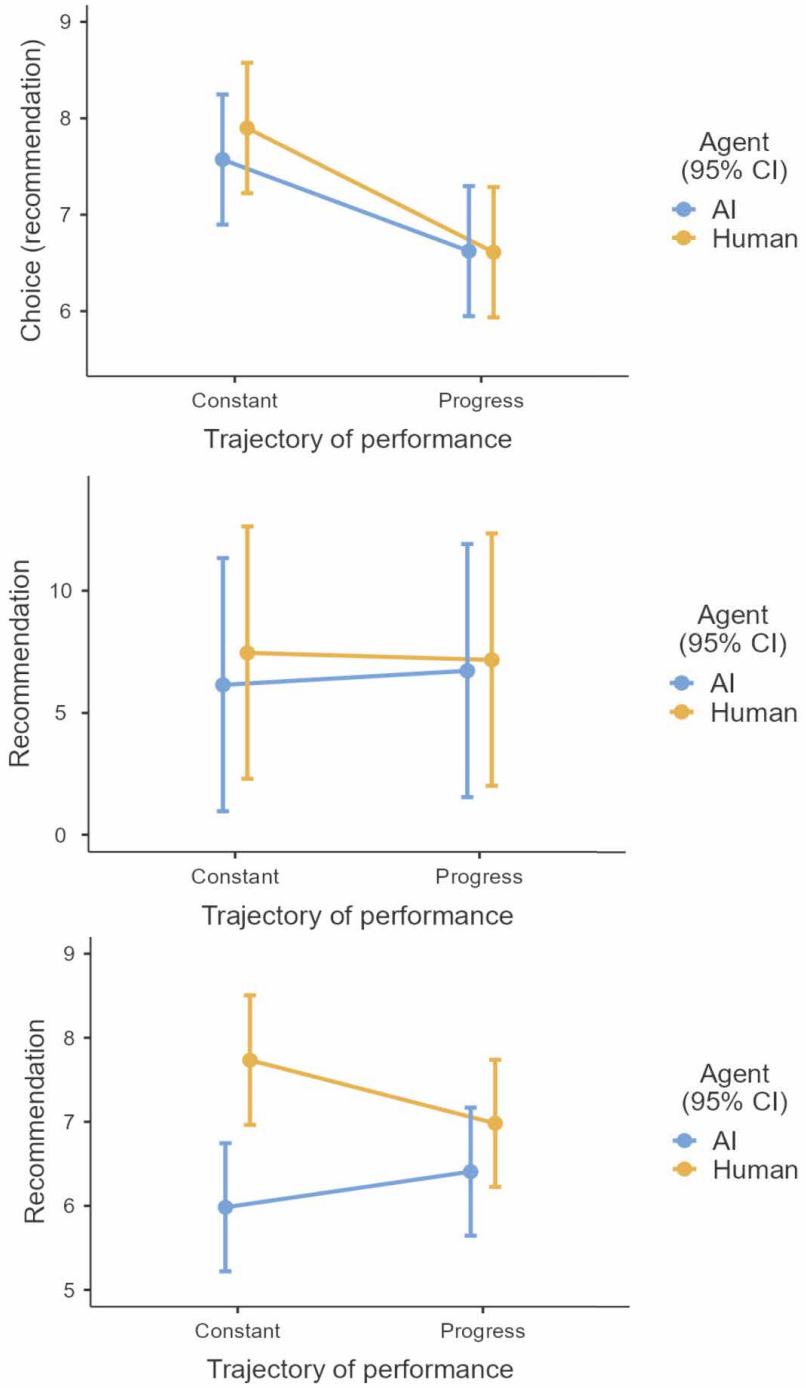
The significant main effect of agent type, with humans being rated higher overall than AI, aligns with previous research showing a general preference for human performance over AI performance in creative domains (Magni *et al*., 2024).

The significant interaction between agent type and trajectory of performance provides important insights into potential strategies for mitigating AI aversion. The preference for human agents was especially large in the constant performance condition, suggesting that when performance shows no improvement, people are more likely to trust human consistency than AI consistency. However, this preference reduced when there was improvement in

performance, indicating that demonstrating an AI's capacity to learn and improve might be an effective way to increase its acceptance in creative fields.

Crucially, we noticed that in all of the experiments in which we used the recommendation scores as a dependent variable, there was a negative effect of the agent (human – AI) x trajectory of performance (progress – constant) interaction (see the comparison in Figure 7). This suggested that the effect was robust. To study it more closely and to combine our findings across experiments, we conducted an internal meta-analysis.

FIGURE 7. The Comparison of Agent x Trajectory of Performance Interactions from Experiments 1, 2, and 4



Source: own elaboration.

 DOI: 10.7206/cid.3071-7973.9

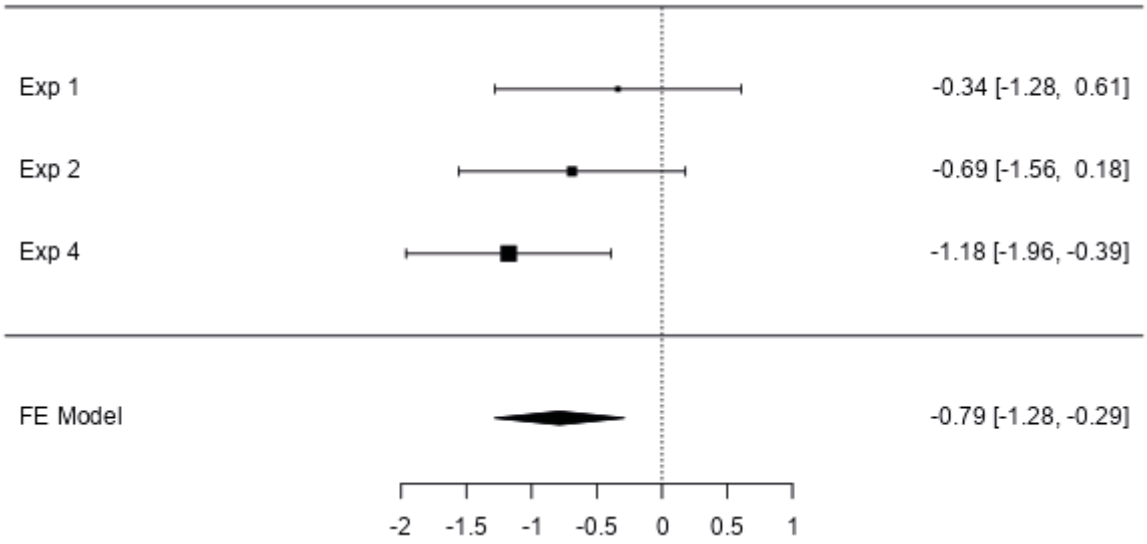
Internal Meta-analysis

We conducted an internal meta-analysis to examine the robustness and magnitude of the interaction effect between agent type (human vs AI) and trajectory of performance (progress vs constant). This analysis allowed us to combine our findings and draw more confident conclusions about the overall patterns in our data.

We focused on the effects of models in three experiments (Experiments 1, 2, and 4) that used recommendation as the dependent variable. We combined the estimated effects using the MAJOR package for JAMOVI (https://github.com/kylehamilton/MAJOR.git), using the recommended Restricted Maximum-Likelihood estimator model with a step length of 0.5 for the Fisher algorithm adjustment.

Results

FIGURE 8. Forest Plot of the Results of the Internal Meta-analysis: Estimated Effect Size for the Agent x Trajectory of Performance Interaction



Note. FE Model stands for Fixed Effects model.

Source: own elaboration.

Figure 8 presents the forest plot of our meta-analysis. Because there was no heterogeneity Q(3)=1.86, p=.395, we used a Fixed Effects model. We saw a significant overall effect size of -0.79 [-1.28, -0.29] for the interaction between agent type and performance trajectory. This negative effect suggests that the preference for human agents over AI was significantly reduced when there was progress in performance.

Crucially, the direction of this effect was consistent across all three experiments included in the meta-analysis, despite their methodological differences.

Discussion

The results of our meta-analysis provide strong support for a robust interaction effect between agent type and performance trajectory. The negative effect size shows that showing progress in performance significantly reduces the preference for human agents over AI. This finding has important implications for our understanding of algorithm aversion and appreciation,

as well as for strategies to increase acceptance of AI systems in various domains, including creative fields.

Despite differences in methodology and context, the consistency of this effect across our experiments further strengthens the finding. It suggests we may generalize the tendency for improvement to mitigate preferences for human agents across different presentation formats and domains.

These results highlight the importance of showing improvement in performance in AI systems as a potential way of increasing their acceptance and reducing algorithm aversion. The moderate to large effect size suggests that this factor plays an important role in how people judge AI performance concerning human performance.

General Discussion

Our series of experiments provides insights into how people evaluate the performance of AI systems and human artists in the domain of tattoo design. Through four experiments, we studied the interplay between performance trajectory (consistent vs. improving), agent type (human vs. AI), and domain (creative field: tattoos vs. non-creative fields). We wanted to contribute to the ongoing discussion on AI acceptance in creative domains. The results of our experiments partly align and partly differ from the previous studies on performance perception and AI judgments.

One of our findings was the consistent preference for constant performance in humans within the tattoo design domain. This result contrasts with previous studies (Soliman & Buehler, 2018a; 2018b), which generally found that people prefer improvement in performance when evaluating candidates for jobs. Our findings suggest that creative endeavors, or at least tattoo design, may be a special case compared to measures of objective performance in other domains that we utilized (e.g., sales or logistics). This difference became particularly visible when we introduced non-creative domains into our model in Experiment 2. The effect of preferring constant performance ceased to exist in those other domains, further showing that the creative domain might be a special case in terms of evaluation (Magni *et al*., 2024).

We investigated whether how people perceive tattoo design as a talent-based vs. skill-based domain might explain the preference for constant performance. However, our results in Experiment 3 showed no significant differences in how participants perceived the source of performance (innate talent vs. learned skill) between creative (tattoo design) and non-creative (law/logistics/sales) domains. This suggests that we cannot explain the preference for constant performance in tattoo design by the perception of the field as more talent-based than other domains.

Because of the incongruencies, we also examined whether the mode of presentation (pictures of tattoos vs. graphs of performance) might influence participants' evaluations. Contrary to our expectations, Experiment 4 showed that the format in which performance was presented did not have a significant main effect on recommendations. There was also no two-way interaction between the mode of presentation and performance trajectory. This result suggests that the discrepancies we saw between our earlier experiments were not due to differences in presentation format.

 DOI: 10.7206/cid.3071-7973.9

Perhaps the lack of preference for improvement over constant performance we observed is task-specific. People might see tattoo design as a special type of creative task. We also only studied it in the context of making hiring decisions. For future directions in research regarding the differences in Human vs AI performances, we would suggest exploring more types of tasks in the creative field, for example, paintings (Colton, 2012), composing music (Du Sautoy, 2019), or writing poems (Gervás, 2019). This would allow for greater generalizability of our findings (Yarkoni, 2022).

Moreover, the type of decisions that participants made may have mattered. We only tested future-oriented decisions (i.e., hiring decision context), as opposed to the past-oriented decision type (i.e., rewarding a person for their past performance) described in the literature (Soliman & Buehler, 2018a; 2018b). The untested decision type could have an effect contrasting with our findings. The exact trajectory of the performance in our studies may have also influenced the result. In the progress condition, we always showed performance that started below and ended at the exact level of the corresponding constant score. For example, we would put together a constant "7" out of "10" performance with a progress performance going from "4" to "7" out of "10." In real life, evaluators usually choose between many candidates; some show constant performance, some show progress, and some even regress. Furthermore, some candidates who improve already start from the level of performance that is shown by consistently performing candidates and might greatly exceed it (Williams & Gilovich, 2012). One size does not fit all.

We also have to consider the influence of effort heuristics (Kruger *et al*., 2004). We presented the performance as pictures and graphs, and following Soliman and Buehler (2018a; 2018b), we assumed that the participants saw improving performance as the result of the candidate's effort. However, we did not directly ask about it, and therefore, we cannot say much about how "effortful" progressing AI seemed in the eyes of our participants.

We have to bear in mind that, as we conducted the research over three years, the AI attitudes of our participants from Studies 1–4 could have differed. Experiment 1 coincided with the release of the breakthrough language model Chat-GPT by OpenAI. Since then, people might have become more acquainted with the technology, and their attitudes towards it might have changed (Fast & Horvitz, 2017). Still, our results show rather similar attitudes toward the recommendations of AI models in hiring decisions across the years, despite moral controversies surrounding AI-generated content (Shoemaker, 2024).

So what can we make of our results? We found an interesting interaction regarding the agent (human vs AI) and trajectory of performance (constant vs progress) that prevailed across the experiments. Humans are much preferred over AI in the constant performance condition, but in the progress condition, there is barely any difference between the two. This might suggest that even though there exists some kind of algorithm aversion mechanism in the tattoo-design domain when both of the agents are comparably constant, it disappears when both of the agents show comparable progress. To make sure that the effect persists, we conducted an internal meta-analysis of the experiments, which showed that highlighting AI's capacity to learn and improve may reduce AI aversion, encouraging its adoption in high-performing tasks. Wider use of algorithms could lead to significant efficiency gains and better decision-making across various fields. Understanding why people object to it is crucial for promoting innovation and effective human-AI collaboration.

Supporting Information

In the Supplement, we attached the first experiment that we conducted on the topic. However, we acknowledged that it has potential problems associated with the confusing way we presented materials to our participants, and thus, we removed it from the main part of the paper. Moreover, we attach material validation checks from Experiment 1.

Data availability statement

Data and materials for all of the experiments are available at:
https://osf.io/uvdpx/?view_only=035152b3b92f4a1b83ed88b3c26a051d.
Experiment 2 was pre-registered: https://aspredicted.org/5G5_VTN.
Experiment 3 was pre-registered: https://aspredicted.org/Z67_TRN.
Experiment 4 was pre-registered: https://aspredicted.org/G84_XYR.

Funding statement

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by grant no. 2020/38/E/HS6/00282 from the National Science Centre (NCN, Poland) to Michał Białek. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

IRB – ethics committee statement

All of the studies were accepted by the Ethics Committee of the Institute of Psychology at the University of Wrocław.

Conflict of Interest

We have nothing to declare.

Acknowledgements

We would like to sincerely thank Anna Wróblewska for her help with designing materials for Experiment 2.

References

Aczel, B. (2024). Let the data talk: Embrace exploratory research. *Nature*, *635*(8040), 788–788. https://doi.org/10.1038/d41586-024-03826-z

Borkowska, A., & Milczarski, W. (2021). *Efekt SNARC. Przegląd aktualnych badań.* In: J. Kowal & K. Chatzipentidis (eds.), *Nowe problemy psychologii. Przegląd zagadnień*. (Vol. 18/19, pp. 198–213). Wrocław: Uniwersytet Wrocławski.

Colton, S. (2012). The Painting Fool: Stories from Building an Automated Painter. In: J. McCormack & M. d'Inverno (eds.), *Computers and Creativity* (pp. 3–38). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31727-9_1

Dietvorst, B.J., Simmons, J.P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Du Sautoy, M. (2019). Can AI ever be truly creative? *New Scientist*, *242*(3229), 38–41. https://doi.org/10.1016/S0262-4079(19)30840-1

Fast, E., & Horvitz, E. (2017). Long-Term Trends in the Public Perception of Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1), 963–969. https://doi.org/10.1609/aaai.v31i1.10635

Gallucci, M. (2019). GAMLj: *General analyses for linear models.* [*jamovi module*]. Available from: https://gamlj.github.io/.

Gervás, P. (2019). Exploring Quantitative Evaluations of the Creativity of Automatic Poets. In T. Veale & F.A. Cardoso (Eds.), *Computational Creativity* (pp. 275–304). Springer International Publishing. https://doi.org/10.1007/978-3-319-43610-4_13

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Kruger, J., Wirtz, D., Van Boven, L., & Altermatt, T.W. (2004). The effort heuristic. *Journal of Experimental Social Psychology*, *40*(1), 91–98. https://doi.org/10.1016/S0022-1031(03)00065-9

Kuncel, N. R. (2008). Some New (and Old) Suggestions for Improving Personnel Selection. *Industrial and Organizational Psychology*, *1*(3), 343–346. https://doi.org/10.1111/j.1754-9434.2008.00059.x

Magni, F., Park, J., & Chao, M.M. (2024). Humans as Creativity Gatekeepers: Are We Biased Against AI Creativity? *Journal of Business and Psychology*, *39*(3), 643–656. https://doi.org/10.1007/s10869-023-09910-x

Patro, K., & Haman, M. (2012). The spatial–numerical congruity effect in preschoolers. *Journal of Experimental Child Psychology*, *111*(3), 534–542. https://doi.org/10.1016/j.jecp.2011.09.006

Russell, S.J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (Fourth Edition). Pearson.

Sawicki, P., & Białek, M. (2017). Overlapping defaults. The case of intertemporal choices. *Polish Psychological Bulletin*, *48*(4), 440–444. https://doi.org/10.1515/ppb-2017-0050

Shoemaker, E. (2024). Is AI Art Theft? The Moral Foundations of Copyright Law in the Context of AI Image Generation. *Philosophy & Technology*, *37*(3), 1–21. https://doi.org/10.1007/s13347-024-00797-x

Soliman, M., & Buehler, R. (2018a). Evaluating performance over time: Is improving better than being consistently good? *The Journal of Social Psychology*, *158*(3), 271–284. https://doi.org/10.1080/00224545.2017.1341373

Soliman, M., & Buehler, R. (2018b). Why Improvement Can Trump Consistent Strong Performance: The Role of Effort Perceptions. *Journal of Behavioral Decision Making*, *31*(1), 52–64. https://doi.org/10.1002/bdm.2039

The jamovi project (2021). jamovi. (Version 2.2) [Computer Software]. Retrieved from https://www.jamovi.org.

Williams, E.F., & Gilovich, T. (2012). The better-than-my-average effect: The relative impact of peak and average performances in assessments of the self and others. *Journal of Experimental Social Psychology*, *48*(2), 556–561. https://doi.org/10.1016/j.jesp.2011.11.010

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences, 45,* e1. https://doi.org/10.1017/S0140525X20001685