Journal of Experimental Psychology: General

The Artificial Intelligence Disclosure Penalty: Humans Persistently Devalue AI-Generated Creative Writing

Manav Raj, Justin M. Berg, and Rob Seamans

Online First Publication, January 8, 2026. https://dx.doi.org/10.1037/xge0001889

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

https://doi.org/10.1037/xge0001889

The Artificial Intelligence Disclosure Penalty: Humans Persistently Devalue AI-Generated Creative Writing

Manav Raj[1], Justin M. Berg[2], and Rob Seamans[3]

[1] The Wharton School, University of Pennsylvania
[2] Ross School of Business, University of Michigan
[3] Stern School of Business, New York University

Although preliminary evidence suggests that humans often react aversely to artificial intelligence (AI)-generated creative works, we have little understanding of how robust or persistent these reactions may be. In a series of 16 preregistered experiments ($N = 27{,}491$), we examine how evaluations of creative writing are affected by whether participants believe the content is produced with an AI model. We find consistent evidence of an AI disclosure penalty: Participant evaluations of creative writing decrease when they believe writing samples were written by an AI model—or with the help of one—rather than a human author alone, and this effect is mediated by perceived authenticity. The AI disclosure penalty is sticky, persisting across evaluation metrics, contexts, kinds of written content, and multiple interventions derived from prior research aimed at moderating the effect. Our results indicate that AI disclosure penalties about creative writing may be stubbornly difficult to mitigate, at least at this time.

Public Significance Statement

Across 16 experiments with over 27,000 participants, we show that people tend to evaluate creative writing less favorably when they believe it was written by artificial intelligence (AI), because they see it as less authentic than if a human author had produced the exact same work without AI. This bias is persistent and difficult to reduce, even when using techniques that mitigate aversion to AI in other contexts.

Keywords: artificial intelligence, creativity, algorithmic aversion, idea evaluation

Supplemental materials: https://doi.org/10.1037/xge0001889.supp

Recent advances have changed the scope of tasks that artificial intelligence (AI) tools can accomplish. Advanced large language models (LLMs) are capable of high levels of creativity (Hubert et al., 2024) and can produce high-quality, humanlike creative content, including creative writing, visual arts, and music (e.g., Chow, 2023; Doshi & Hauser, 2024; Zhou & Lee, 2024). AI-generated creative goods have found critical and audience acclaim in some domains, as an AI-generated digital art piece won the Colorado state art

Aysecan Boduroglu served as action editor.

Manav Raj https://orcid.org/0000-0002-3657-7843

All errors are the authors' own. The authors have shared preliminary versions of this article on online preprint websites (SSRN and arXiv) and have posted the data, code, and past versions of the article on the Open Science Framework at https://osf.io/6tybe. The authors have also presented earlier versions of the draft at the New York University Stern Artificial Intelligence (AI) Workshop and the Seminar for Experimental Organizational Science. Manav Raj received funding from the Wharton Behavioral Lab. Justin M. Berg received funding from the Stanford Graduate School of Business. The authors thank the Wharton Behavioral Lab and the Stanford Graduate School of Business for funding assistance and thank the New York University Stern AI Workshop and the Seminar for Experimental Organizational Science for the feedback.

Manav Raj: Conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, visualization, writing–original draft, and writing–review and editing. Justin M. Berg: Conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, resources, supervision, writing–original draft, and writing–review and editing. Rob Seamans: Conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, writing–original draft, and writing–review and editing.

Manav Raj played an equal role in conceptualization, funding acquisition, investigation, methodology, project administration, resources, writing–original draft, and writing–review and editing. Justin M. Berg played an equal role in conceptualization, funding acquisition, investigation, methodology, project administration, resources, writing–original draft, and writing–review and editing. Rob Seamans played an equal role in conceptualization, funding acquisition, investigation, methodology, project administration, resources, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Manav Raj, The Wharton School, University of Pennsylvania, 3620 Locust Walk, Philadelphia, PA 19104, United States. Email: manavraj@wharton.upenn.edu

competition (Roose, 2022) and an AI-generated song mimicking the artists Drake and The Weeknd quickly accumulated millions of streams across TikTok, Spotify, and YouTube (Coscarelli, 2023). At the same time, producers and consumers of creative goods have responded to the emergence of AI-generated content with angst, suggesting that this heralds the "death of artistry" (Roose, 2022) and that AI-generated creative goods are "the opposite of art" (Shaffi, 2023). Fundamentally, these rising tensions raise questions regarding whether and to what extent a "human" element carries value in the eyes of consumers of creative content and whether humans can appreciate creative content when they know it was generated by AI (e.g., Gonzalez, 2023; Mineo, 2023; Ornes, 2019; Shank, 2025).

Recent studies have sought to shed light on such questions, and some key findings have emerged from this nascent but growing body of research. One consistent finding is that when people are not told whether creative goods are generated by AI, they are often unable to distinguish between AI- and human-generated creative goods, suggesting that content generated by AI is not inherently different from content generated by human creators (e.g., Hitsuwari et al., 2023; Köbis & Mossink, 2021). However, an emerging body of literature also shows that when people are explicitly told that a creative good was generated by AI, they tend to appreciate it less than if they are told the very same good was generated by a human (e.g., Bellaiche et al., 2023; Horton et al., 2023; Jago, 2019; Shank, 2025; Shank et al., 2023; Wu et al., 2020). We call this phenomenon the *AI disclosure penalty*—the reduction in appreciation of a creative good that occurs when consumers are informed that the good was created by or with the assistance of AI. To date, the AI disclosure penalty has been documented with respect to consumers' evaluations of artistic images (Horton et al., 2023), paintings (Bellaiche et al., 2023; Jago, 2019; Wu et al., 2020), songs (Jago, 2019; Shank et al., 2023), and poetry (Wu et al., 2020). These studies show that audiences react negatively when they are informed that AI tools were involved in the production of creative goods—they appreciate the exact same creative goods more when they think they were created by humans without AI involvement.

This emerging body of research suggests that consumers stand to garner enjoyment and value from AI-generated creative goods if only they remain unaware that AI was involved in the creation process. In this way, AI disclosure presents a thorny dilemma: Creators cannot be transparent about using AI in producing creative content without undermining appreciation for that very content—unless there are ways to reduce AI disclosure penalties. However, we currently have little understanding of what conditions, if any, may mitigate or exacerbate AI disclosure penalties. The relatively small number of studies looking at moderators of AI disclosure penalties has examined whether framing the disclosure as human–AI collaboration, as opposed to purely AI-generated, reduces the penalty. In one article focusing on evaluations of artistic images (Horton et al., 2023), the authors found that disclosing human–AI collaboration reduced the AI disclosure penalty compared with purely AI, but purely human was appreciated more than human–AI collaboration. Another article (Tigre Moura et al., 2023) found this same pattern of results for tangible products (artistic images), but not for relatively intangible products (songs). Given the sparse and mixed results on moderators of AI disclosure penalties, existing research says relatively little about whether or when AI disclosure penalties can be mitigated.

Furthermore, prior research has not thoroughly explored how this phenomenon may extend to creative writing—even though textual output is what the increasingly popular and sophisticated AI tools such as OpenAI's ChatGPT, Anthropic's Claude, Google's Gemini, and other LLMs excel at generating. While AI disclosure penalties have been documented with respect to poetry generated by AI models before the advent of LLMs (e.g., Wu et al., 2020), there is less evidence regarding the presence of such penalties for other forms of creative writing or with creative writing produced by newer, more sophisticated AI tools. Other studies that consider the effects of AI involvement or disclosure on creative writing also do not fill this void. While there is evidence that humans may struggle to distinguish AI- versus human-generated creative writing and may discount the likelihood that high-quality creative goods are produced by AI (e.g., Hitsuwari et al., 2023; Köbis & Mossink, 2021), these studies say little about whether aversion to AI-generated creative writing is driven by the content, AI disclosure, or some combination of both. Given that the use of AI tools to assist with (or completely automate) creative writing work is already widespread (with and without disclosure) and will likely only increase in the years to come (e.g., Rogers, 2024; Sherrer, 2025; Ticong, 2025), understanding the relationship between AI disclosure and evaluations of creative writing has practical and ethical implications as we consider how producers of written content use and disclose the use of AI.

In the present research, we provide large sample evidence of how the disclosure of the use of AI in the generation of creative written content affects human evaluation and probe for heterogeneity in this effect. In a series of 16 preregistered experiments (N = 27,491) conducted between March 2023 and June 2024, we examine the effect of AI disclosure on evaluations of creative writing and what conditions (if any) may moderate this effect. We also consider what potential mechanisms may mediate the relationship between AI disclosure and evaluations of creative writing. We conclude with meta-analyses summarizing the results of the 16 studies. Across our studies and meta-analyses, we document consistent evidence of an AI disclosure penalty: Participant evaluations of creative writing samples decrease when they believe that the writing samples were written by, or with the help of, an AI model rather than a human author without the use of AI. This AI disclosure penalty is mediated by perceived authenticity, suggesting that at least at this time in history, people tend to see AI-generated creative writing as relatively inauthentic and therefore less worthy of their appreciation.

Importantly, this AI disclosure penalty is remarkably persistent, holding across the time period of our study; across different evaluation metrics, contexts, and kinds of written content; and across interventions derived from prior research aimed at moderating the effect. Specifically, we designed and tested a variety of interventions with the intention to moderate the AI disclosure penalty. Our empirical approach was open ended and exploratory in nature, and the direction of our inquiry unfolded iteratively as we ran studies. However, throughout the process, we sought to moderate the AI disclosure effect by building upon prior literature on algorithmic aversion, specifically work showing that reactions to AI can differ in hedonic versus utilitarian domains (Longoni & Cian, 2022), when an AI tool is more versus less humanized (e.g., Burton et al., 2020; Schanke et al., 2021), when participants' perceptions of an AI's capabilities are higher versus lower (Bellaiche et al., 2023), and when a human is "in the loop" (e.g., Dietvorst et al., 2018;

Hong et al., 2022; Horton et al., 2023). Despite our deliberate attempts, we were unable to find consistent moderation of the effect, suggesting that the AI disclosure penalty with respect to creative writing is, at least at the time of our study, quite robust and surprisingly difficult to mitigate.

Beyond the timeliness and practical implications of this work, this research also addresses foundational psychological questions that span multiple subfields, including cognitive psychology, social psychology, and consumer behavior. This research contributes to cognitive psychology by deepening our understanding of the implications of individuals' assessment of who or what possesses mental life (e.g., Gray et al., 2007). It informs social psychology by shedding light on how dimensions of identity held by a creator (i.e., human vs. nonhuman) shape perceivers' appraisals of the content they produce; this particular dimension of identity is not commonly explored in prior research, which tends to focus on dimensions of identity within the broader "human" category (e.g., divisions based on race, gender, socioeconomic status; Cuddy et al., 2008). Finally, our research also has implications for consumer psychology and judgment and decision-making research by illustrating how disclosure effects and algorithmic aversion influence preferences and evaluations, particularly in creative domains (e.g., Dietvorst et al., 2015; Longoni & Cian, 2022). By integrating these perspectives, our research advances broader theories of human perception, social cognition, and the psychological mechanisms underlying resistance to AI.

Method and Results

Overview of Studies

Our aim in this research was to determine whether AI disclosure affects evaluations of creative writing and to identify when, and under what conditions, that effect may grow stronger or weaker. To do so, we conducted a series of 16 preregistered experiments. We let the results from each study guide the next, taking an iterative, exploratory approach grounded in theories of algorithmic aversion. Across the 16 experiments, we found consistent evidence that disclosing the use of AI lowered readers' evaluations of creative writing. Despite our deliberate attempts to moderate this AI disclosure penalty through various interventions derived from prior research, the negative effect persisted across studies, demonstrating a surprising level of robustness. For clarity, we group the 16 experiments into five clusters (summarized in Table 1) that trace how our inquiry unfolded.

First, in Studies 1a–1e, we tested whether the impact of AI disclosure depends on content features that make the writing feel more or less human, such as narrative perspective, format, emotional tone, or the humanness of characters (Castelo et al., 2019). AI disclosure led to worse evaluations, and any moderation by these content features was small and inconsistent. Second, drawing on evidence that algorithmic aversion is stronger for hedonic than utilitarian judgments (Longoni & Cian, 2022), Study 2 compared artistic and nonartistic evaluation contexts. Reframing the context in this way did not soften the AI disclosure penalty; the penalty was equally strong in both settings. Third, motivated by work showing that altering perceptions of AI can curb resistance (Burton et al., 2020; Diel et al., 2021; Schanke et al., 2021), Studies 3a–3e manipulated information about the AI's capabilities and the extent to which it was anthropomorphized, neither of which reliably reduced the AI disclosure penalty. Fourth, because people may hold mixed feelings toward AI, Studies 4a and 4b examined whether disclosure increases ambivalence or simultaneous positive and negative evaluations (Thompson et al., 1995). Instead of heightening ambivalence, disclosure simply made judgments more negative, consistent with our prior studies. Finally, Studies 5a–5c tested a human-in-the-loop framing in which stories were labeled as products of human–AI collaboration, an approach shown to reduce algorithmic aversion (Dietvorst et al., 2015; Horton et al., 2023; Mellamphy, 2021). Collaboration labels offered no relief; readers still discounted the writing once AI involvement was disclosed. Across studies, we collected data on a variety of mechanism measures motivated by each study, and we tested whether the relationship between AI disclosure and evaluation is mediated by each potential mechanism measure (we report these mediation results for all studies after the main results are reported for each study). Across studies, we found consistent support that reduced perceptions of authenticity mediate the AI disclosure penalty.

Following the 16 experiments, we report a meta-analysis summarizing our findings and effect sizes. Overall, our studies and meta-analyses show a robust, authenticity-driven bias against AI-generated (or AI-assisted) creative writing that is resistant to several evidence-based interventions that have worked on similar biases in the past.

Transparency and Openness

Studies were approved by the institutional review board at the corresponding author's home institution with the Identification No. 852975. All analyses were conducted using the statistical software programming tool STATA (StataCorp., 2021). In addition to using built-in STATA programming commands, we utilized the following user-generated STATA commands to conduct our analyses and/or organize our results: parmest, reghdfe, ftools, and asdoc (Correia, 2023a, 2023b; Newson, 2022; Shah, 2021). Data and code necessary to reproduce the findings, as well as access to study materials and an online Supplemental Appendix (which includes tables with the full models, summary statistics of dependent variables by condition, correlation matrices, and Cronbach's αs for scale measures), are available at https://osf.io/6tybe. This project meets the *Journal of Experimental Psychology: General* standards for transparency and openness across all seven of the Transparency and Openness Promotion guidelines. We cite all data, program code, and methods developed by others, we provide access to the raw data on which the study conclusions are based, we describe how to access code needed to reproduce analyses, we provide access to study materials in a trusted repository, we transparently discuss the study design and analysis plan, and we report information regarding the preregistration of each study along with a link to access each preregistration.

Across studies, we targeted sample sizes of approximately 100 participants per experimental cell. While we did not run a formal power analysis, we chose this benchmark as we believed it would be large enough to capture potentially small effect sizes and to explore heterogeneity in the effect within noisy online experimental contexts. In collecting demographic information in each study, we asked participants to self-report information among a set of options about their gender (male, female, nonbinary/third gender, prefer to self-describe, and prefer not to say for

Table 1
Summary of Study Designs and Key Results

Study	Focus of study	Key result	Creative writing sample used (ChatGPT-generated unless noted)	Date	*n*
Studies 1a–1e: How do content characteristics shape AI disclosure effects?					
1a	Test whether AI disclosure effects depend on perspective (first vs. third person) in poetry	AI disclosure lowered enjoyment, creativity, and quality; first-person penalty significantly larger for enjoyment, marginally for creativity, and *ns* for quality.	Six free-verse poems on impermanence, memory, regret (each in first and third person)	March 1, 2023	1,342
1b	Examine if format (poetry vs. prose) and emotional tone (positive/neutral/negative) moderate AI disclosure effects	AI disclosure lowered enjoyment; no robust moderators.	Six pieces; poem and matching prose on love, exploring Los Angeles, depression	July 12, 2023	2,755
1c[a]	Replicate 1a with higher quality GPT-4 prose after mixed results.	Strong AI disclosure penalty on all metrics; no robust moderation (just marginally larger for first-person enjoyment)	Six prose pieces on visiting grandmother, town fair, firefighter for a day (first vs. third person)	August 24, 2023	2,753
1d[a]	Vary character type (human vs. animal vs. alien) to see if the humanness of content shapes the penalty	Replicated baseline penalty; smaller when characters were aliens; no change with animals	Six third-person prose pieces on grandmother visit and trip to New York City (with human, animal, or alien protagonists)	September 18, 2023	2,739
1e[a]	Refine 1d by adding a robot condition and minimizing unintended anthropomorphism (make robots, animals, and aliens more clearly not human)	Penalty replicated; this time larger with animal characters; marginally smaller with robots	Eight third-person prose pieces on elder visit and trip to New York City (with human, animal, alien, or robot protagonists)	October 13, 2023	3,510
Study 2: How does evaluation context shape AI disclosure effects?					
2	Test whether framing the task as utilitarian (objective quality) versus hedonic (artistic taste) weakens the penalty	Penalty persisted; no moderation by context.	Three first-person prose pieces (grandma's garden, firefighter for a day, town fair)	November 14, 2023	3,590
Studies 3a–3e: How does perception of AI shape AI disclosure effects?					
3a	Assess whether boosting beliefs about AI's emotional or cognitive capabilities reduces the penalty	Penalty remained robust; manipulations convinced participants of intended capabilities but did not moderate the penalty.	Three first-person prose pieces (same as Study 2)	December 19, 2023	1,526
3b	Provide objective, humanizing, or AI-as-tool, info to see if extra detail tempers aversion	Penalty remained; "tool" label statistically offset penalty but itself depressed ratings, yielding no net gain.	Three first-person prose pieces on predawn running, embarrassed by dad, overheard conversation	January 23, 2024	1,799
3c	Test whether anthropomorphizing the AI (name, gender, backstory) reduces the penalty	Baseline penalty *ns*; humanizing the AI backfired and produced a penalty	Three first-person prose pieces on dad, romantic partner, friend	February 14, 2024	864

(*table continues*)

Table 1 (*continued*)

Study	Focus of study	Key result	Creative writing sample used (ChatGPT-generated unless noted)	Date	*n*
3d[a]	Replicate 3c and test authenticity versus identity-threat mediation	Baseline penalty returned; humanization no longer moderated; authenticity mediated and identity threat did not	Three first-person prose pieces on grief after loss, coffee-shop reflection, bond with dog	March 5, 2024	909
3e[a]	Second replication with even stronger humanization language to resolve inconsistencies between 3c and 3d	Penalty again significant; no interaction with stronger humanization	Three first-person prose pieces on first day at new college, dad story, relationship reflection	March 19, 2024	908
		Studies 4a and 4b: How does AI disclosure affect ambivalence about the content?			
4a	Investigate whether disclosure increases ambivalence toward the content	Penalty persisted; no effect on ambivalence	Three first-person prose pieces on dad, romantic partner, friend	April 8, 2024	423
4b	Test if humanizing AI heightens disclosure-induced ambivalence	Penalty remained robust; no effects for ambivalence	Three first-person prose pieces on dad, romantic partner, friend	April 12, 2024	1,280
		Studies 5a–5c: How does a human-in-the-loop framing shape AI disclosure effects?			
5a[a]	Test a “human-in-the-loop” framing (AI described as a tool under human control)	Penalty stable; tool framing did not moderate (and had its own negative main effect, like Study 3b).	Three first-person prose pieces on dad, romantic partner, friend	April 29, 2024	856
5b[a]	Compare disclosure of AI-only versus human-only versus human–AI collaboration	Penalty persisted and was the same for AI-only and human–AI collaboration.	Three first-person prose pieces on dad, romantic partner, friend	May 3, 2024	665
5c[a]	Replicate 5b with award-winning stories written by human authors	Same pattern as 5b: Penalty persisted for AI-only and human–AI collaboration.	Five award-winning human-written short stories	June 6, 2024	1,572
				Total *N*	27,491

Note. This table provides a high-level overview of all studies run. The writing sample column describes the stimuli used in each study, the date column indicates the date of data collection, and *n* indicates the final sample size for the analyses. AI = artificial intelligence.

[a] Study wherein we collect data on perceived authenticity and test for mediation of the AI disclosure penalty.

some experiments), race (White or Caucasian, Black or African American, American Indian/Native American or Alaska Native, Asian, Native Hawaiian or other Pacific Islander, other, or prefer not to say; participants could select multiple), and age (18–99). All studies included two attention check questions, and participants who failed either attention check were dropped from the analysis.

Studies 1a–1e: How Do Content Characteristics Shape AI Disclosure Effects?

In our first five studies (Studies 1a–1e), we explored whether and to what extent evaluations of AI-generated creative writing may differ depending on characteristics of the content itself. Building on prior literature that suggests that algorithmic aversion tends to be greater when AI is engaging in more humanlike activities (Castelo et al., 2019), we examined the moderating role of content characteristics that are likely to affect the humanness of the creative writing. Specifically, we considered whether AI disclosure effects differ for the following content characteristics: perspective (first vs. third person, Studies 1a and 1c), format (poetry vs. prose, Study 1b), emotionality (Study 1b), or humanness of the characters (Studies 1d and 1e).

Across studies, we asked participants to read and evaluate AI-generated writing samples, created using ChatGPT. We chose to use ChatGPT because, at the time of the initial study in March 2023, it was the most well-known LLM. For consistency, we then used ChatGPT for all subsequent studies. Participants were informed that they would be taking part in a study to examine how people evaluate written content. In each study, all participants were asked to read and evaluate a writing sample. We randomly assigned participants to either be informed that the writing sample was written by "the artificial intelligence (AI) tool ChatGPT" or by "a poet" (Study 1a) or "a participant in a prior study" (Studies 1b–1e). After reading the writing samples, participants were asked to evaluate the quality, creativity, and enjoyment of the work on a 1–7 scale. We collected further data on demographic information, a 20-item scale of the Big 5 personality traits (Donnellan et al., 2006), and data on participant familiarity with ChatGPT[1] and similar AI tools.

In Studies 1c–1e, we also collected data on a variety of measures to explore potential mechanisms of the AI disclosure penalty. We took a broad approach, collecting a number of different potential mediators and then testing them in an open-ended manner. The logic underlying these mediators was to choose items that we thought were linked to the human element of creative content and measures that prior research on AI or automation suggested may be shaped by the use of AI. We adapted a measure of perceived humanness of the writing sample from Martin and Mason (2022). We collected measures of perceived profundity, emotional response, story, meaningfulness, effort, time, and worth of the writing sample from Bellaiche et al. (2023). We asked participants to assess the perceived authenticity of the writing sample, and we captured participant engagement with the writing sample using the length of time taken to read the sample, a written explanation of the evaluation, and a comprehension check question. We note that the mechanism measures captured in Studies 1a–1e are single-item measures.

Study 1a

Procedure. In Study 1, we explored whether and to what extent evaluations of AI-generated creative writing may differ depending on the perspective of the creative content (first vs. third person). This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 123538, https://aspredicted.org/7L4_V72). This experiment initially included 1,440 participants in the United States recruited via the Prolific platform, who were each compensated $1.75. After dropping participants who failed either of the two attention checks, the final sample included 1,342 participants (648 reported their gender as male, 668 as female, and 19 as other, and seven declined to report; age = 18–91, $M_{\text{age}} = 44.8$, $SD_{\text{age}} = 15.7$). The sample was recruited in a manner so that it was representative of the U.S. population across age, race, and gender. In total, we generated six poems using ChatGPT that were used in the study (Three Topics × Two Perspectives). Across conditions, participants were randomly assigned to one of the six creative writing samples. As an example of the nature of the writing samples, below is the first stanza of the first-person poem about memory used in the study (note that all study materials are available on the Open Science Framework at https://osf.io/6tybe):

I walk through the garden of my mind

Picking flowers of different colors and scents

Some are fresh and bright, others are wilted and faded

Each one holds a fragment of my past

A moment, a feeling, a person

Results. As preregistered, we explored the effect of AI disclosure using analysis of variance (ANOVA). To do so, we used a full factorial model that included main effects, two-way interactions, and the three-way interaction across our three independent variables (Topic × Perspective × AI Disclosure). Condition variables are all dummy coded for this and all other studies. The ANOVA results reveal that AI disclosure has a significant negative effect on evaluations of enjoyment, $F(1, 1{,}330) = 6.61$, $p = .010$; creativity, $F(1, 1{,}330) = 12.07$, $p < .001$; and quality, $F(1, 1{,}330) = 6.77$, $p < .001$. The interaction between AI disclosure and perspective is significant for evaluations of enjoyment, $F(1, 1{,}330) = 4.33$, $p = .038$; marginal for evaluations of creativity, $F(1, 1{,}330) = 6.61$, $p = .088$; and not significant for evaluations of quality, $F(1, 1{,}330) = 1.52$, $p = .218$. The three-way interaction is not significant across any of the evaluation metrics: $F(2, 1{,}330) = 1.50$, $p = .224$ for enjoyment; $F(2, 1{,}330) = 1.44$, $p = .237$ for creativity; and $F(2, 1{,}330) = 1.27$, $p = .280$ for quality.

Study 1b

Procedure. In Study 1b, building on our findings from Study 1a, we explored how evaluations of AI-generated creative writing

[1] Given that familiarity with AI tools seems like it could reduce the AI disclosure penalty, we tested this variable as a possible moderator of the AI disclosure penalty in each study. Of our 16 studies, we find significant moderation in three cases (Studies 1c, 3a, and 3b); this moderation is positive for Studies 1c and 3a but negative in Study 3b. We also did not see a trend emerge such that moderation grew stronger (or weaker) in later studies. As such, familiarity with AI does not appear to be a reliable moderator of the AI disclosure penalty in this set of studies.

may differ depending on the emotionality (neutral vs. negative vs. positive) and format (poetry vs. prose) of the writing sample. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 138224, https://aspredicted.org/LQT_RYY). The experiment initially included 2,880 participants in the United States recruited via Prolific, who were each compensated $1.00. After dropping participants who failed either of the two attention checks, the final sample included 2,755 participants (1,384 reported their gender as male, 1,328 as female, 33 as nonbinary or third gender, and five as other, and five declined to report; age = 18–90, M_{age} = 35.0, SD_{age} = 13.6). We generated six writing samples using ChatGPT that were used in the study—a matched set of poetry and prose for each of three separate topics (one neutral emotionally, one negative emotionally, and one positive emotionally). Across conditions, participants were randomly assigned to one of the six creative writing samples.

Results. As preregistered, we explored the effect of AI disclosure using ordinary least squares (OLS) regression. Our models used creativity, quality, and enjoyment as dependent variables. As independent variables, we included AI disclosure, topic, writing format, and interactions of AI disclosure and topic and AI disclosure and writing format. Considering creativity as a dependent variable, we do not find a significant effect of AI disclosure (p = .412), nor do we find any significant interaction effects (p = .567 and .438 for positive and negative emotionality, respectively). Considering quality as a dependent variable, we do not find a significant baseline effect of AI disclosure (p = .909) or the interaction between AI disclosure and positive emotionality (p = .103); however, we find a negative and marginal interaction between negative emotionality and AI disclosure (p = .053). Considering enjoyment as a dependent variable, we find a negative and significant baseline effect of AI disclosure (p = .029) and do not find evidence of significant interaction effects (p = .774 and .301 for the interaction between AI disclosure and positive and negative emotionality, respectively).

Study 1c

Procedure. While the results of Study 1b presented some evidence that emotionality or format could moderate the effect of AI disclosure, results were weak and inconsistent. In Study 1c, as in Study 1a, we again considered whether evaluations of AI-generated creative writing may differ depending on perspective of the writing sample. While in Study 1a our stimuli were free-verse poems about abstract constructs such as impermanence, memory, and regret, in Study 1c, we instead used narrative prose writing samples. Further, due to advances in AI technologies between Studies 1 and 3, we utilized an updated version of ChatGPT (GPT-4.0), which allowed us to generate writing samples that we believe were of higher quality. We hypothesized that the penalty for AI disclosure will be larger for writing samples written in the first- versus third-person perspective. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 141733, https://aspredicted.org/PH8_SXG). This experiment initially included 2,880 participants in the United States recruited via Prolific, who were each compensated $1.40. After dropping participants who failed either of the two attention check questions, the final sample included 2,753 participants (1,447 reported their gender as male, 1,238 as female, 52 as nonbinary or third gender, and five as other, and 11 declined to report; age = 18–80, M_{age} = 34.8, SD_{age} = 12.8). We generated six writing samples that were used in the study (Three Topics × Two Formats). Across disclosure conditions, participants were randomly assigned to one of the six creative writing samples.

Results. As preregistered, we explored the effect of AI disclosure using OLS regression. Our models used creativity, quality, and enjoyment as dependent variables. As independent variables, we included AI disclosure, topic, perspective, and interactions of AI disclosure and perspective. Across all dependent variables, we find a significant and negative baseline effect of AI disclosure (p = .005 for creativity and quality; p = .001 for enjoyment). We find a marginal, negative interaction effect between AI disclosure and first-person perspective considering enjoyment as a dependent variable (p = .060), but we do not find significant interaction effects for the other dependent variables (p = .815 and .918 for creativity and quality, respectively). The divergence in these results relative to the results in Study 1a could be due to multiple reasons, as follows: (a) We utilize different kinds of stimuli across studies (free-verse abstract poems in Study 1a vs. narrative prose in Study 1c); (b) we are able to take advantage of higher quality AI models in creating the stimuli in Study 1c due to advances in the technology, and it is possible that AI disclosure effects manifest differently based on content quality; and/or (c) Study 1a was conducted in March 2023, a relatively short period of time after the public launch of ChatGPT, and it is possible that the effect of AI disclosure manifested in a different manner once novelty of the technology wore off.

Study 1d

Procedure. Given mixed evidence regarding the moderating role of perspective in Studies 1a versus 1c, in Study 1d, we continued our exploration of how the humanness of creative works may affect AI disclosure effects by examining whether evaluations of AI-generated creative writing may differ depending on the presence of human versus nonhuman characters. Creative writing that contains human characters may be considered more human. Therefore, this study predicted writing samples with human characters, and thus higher humanness, would receive a larger AI disclosure penalty. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 143840, https://aspredicted.org/RV4_B74). This experiment initially included 2,880 participants in the United States recruited via Prolific, who were each compensated $1.20. After dropping participants who failed either of the two attention checks, the final sample included 2,739 participants (1,319 reported their gender as male, 1,380 as female, 27 as nonbinary or third gender, and two as other, and 11 declined to report; age = 18–83, M_{age} = 37.1, SD_{age} = 13.5).

For two different topics, we generated creative, fictional writing samples that were similar except for manipulating whether the characters were human, animal, or alien. We then pretested writing samples to ensure that the six selected for the experiment were similar in terms of baseline creativity, quality, and enjoyment. As an example of the matched stimuli, we provide the first paragraph of each writing sample for one topic (visiting grandma's house) across our three humanness conditions.

Human condition:

> As the mid-July day began with the sounds of distant lawnmowers and children playing, the gentle sunrise kissed the well-maintained suburban homes in the peaceful neighborhood. Alex made his way to Grandma Hazel's home, located at the end of a peaceful cul-de-sac a ten-minute bike ride from his own. The house flaunted a red-brick

> exterior and a shingled roof which showed the wear of many seasons. More striking than the house itself, however, was the meticulously maintained backyard garden, Grandma Hazel's pride and joy. With an array of flowers, ranging from roses to tulips to hydrangeas, the garden was a vivid spectacle in the quiet suburbia. That day, Alex had been summoned to help her put the finishing touches on the garden for the local "Best Backyard" contest.

Animal condition:

> As the mid-July day began with the sounds of distant lawnmowers and children playing, the gentle sunrise kissed the well-maintained suburban homes in the peaceful neighborhood. Alex, a diligent border collie, made his way to Grandma Hazel's home, located at the end of a peaceful cul-de-sac a ten-minute scamper from his own doghouse. Grandma Hazel, a venerable golden retriever with a regal bearing, lived in a home that flaunted a red-brick exterior and a shingled roof showing the wear of many seasons. More striking than the house itself, however, was the meticulously maintained backyard garden—Grandma Hazel's pride and joy. With an array of flowers, ranging from roses to tulips to hydrangeas, the garden was a vivid spectacle in the quiet suburbia. That day, Alex had been summoned to help her put the finishing touches on the garden for the local "Best Backyard" contest.

Alien condition:

> As the day began with the sounds of distant lawnmowers and children playing, the gentle double-sunrise over the planet Zeebon in the Zorlax Galaxy kissed the well-maintained suburban homes in the peaceful neighborhood. Alex, a young Zorlaxian, made his way to Grandma Hazel's home, located at the end of a peaceful cul-de-sac a ten-minute anti-grav scooter ride from his own. The house flaunted an iridescent exterior and a shingled roof which showed the wear of many seasons. More striking than the house itself, however, was the meticulously maintained backyard garden, Grandma Hazel's pride and joy. With an array of flowers, ranging from Glorbon roses to Zental hydrangeas, the garden was a vivid spectacle in the quiet suburbia. That day, Alex had been summoned to help her put the finishing touches on the garden for the local "Best Backyard" contest.

Results. As preregistered, we explored the effect of AI disclosure using OLS regression. Our models used creativity, quality, and enjoyment as dependent variables. As independent variables, we included AI disclosure, topic, humanness condition, interactions of AI disclosure and the humanness condition, and the interaction between topic and the humanness condition. Across dependent variables, we find a significant and negative baseline effect of AI disclosure ($p < .001$ for creativity and enjoyment; $p = .004$ for quality). Considering creativity and enjoyment, we find a significant and positive interaction between AI disclosure and the "alien" humanness condition, indicating a smaller AI disclosure penalty when the sample contains alien characters ($p = .024$ and $p = .004$, respectively). This effect is nonsignificant considering quality as a dependent variable ($p = .206$). Further, the interaction between AI disclosure and the animal humanness condition is nonsignificant considering creativity, quality, and enjoyment as dependent variables ($p = .737$, .154, and .603, respectively).

Study 1e

Procedure. Study 1e is a constructive replication of Study 1d, further examining whether evaluations of AI-generated creative writing may differ depending on whether the writing sample contains human versus nonhuman characters. We hypothesized that the penalty for AI disclosure will be larger for writing samples containing human (vs. nonhuman) characters. Relative to the prior study, we added a condition with robot characters as well, as we felt that such characters may seem more concordant with the identity of an AI author. Further, we amended the writing samples to mitigate the tendency to anthropomorphize nonhuman characters and thus to ensure that the samples containing human characters were rated higher in humanness. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 147028, https://aspredicted.org/6YN_91V). This experiment initially included 3,840 participants in the United States recruited via the Prolific platform, who were each compensated $1.00. After dropping participants who failed either of the two attention check questions, the final sample included 3,510 participants (1,646 reported their gender as male, 1,794 as female, 56 as nonbinary or third gender, and three as other, and 11 declined to report; age = 18–99, $M_{\text{age}} = 35.6$, $SD_{\text{age}} = 13.0$).

We generated eight writing samples using ChatGPT that were used in this study (Two Topics × Four Humanness conditions). We pretested samples to ensure that the samples containing human characters were perceived as higher in humanness but were rated similarly in terms of quality, creativity, and enjoyment. Across disclosure conditions, participants were randomly assigned to one of the eight creative writing samples.

Results. As preregistered, we explored the effect of AI disclosure using OLS regression. Our models used creativity, quality, and enjoyment as dependent variables. As independent variables, we included AI disclosure, topic, humanness condition, interactions of AI disclosure and the humanness condition, and the interaction between topic and the humanness condition. We find a significant and negative AI disclosure effect considering creativity as a dependent variable ($p = .011$) and enjoyment ($p = .002$); we do not find a significant baseline AI disclosure effect considering quality ($p = .298$). Considering creativity and quality, we find a significant and negative interaction between AI disclosure and the animal humanness condition, indicating a larger AI disclosure penalty when the writing sample contains animal characters ($p = .003$ for creativity; $p < .001$ for quality); we do not find this interaction considering enjoyment as a dependent variable ($p = .742$). However, we do find a marginal and positive interaction between AI disclosure and the "robot" humanness condition ($p = .078$), indicating a marginally smaller AI disclosure penalty on enjoyment for creative writing samples that contain robot characters; this interaction is not significant considering creativity or quality as a dependent variable ($p = .238$ and .375, respectively). We find a marginal negative interaction between the alien humanness condition and AI disclosure considering creativity as a dependent variable ($p = .079$), but this effect is nonsignificant considering quality or enjoyment as a dependent variable ($p = .160$ and $p = .489$, respectively).

Study 2: How Does Evaluation Context Shape AI Disclosure Effects?

While, thus far, we had evaluated whether content characteristics moderate the AI disclosure penalty, we had yet to find consistent evidence of moderation in a manner that felt robust and interpretable. Thus, moving forward, we decided to leverage manipulations that had previously mitigated algorithmic aversion or AI disclosure effects. In Study 2, we probed heterogeneity in AI disclosure effects based on the evaluation context. We considered whether AI disclosure affects the appeal of creative content differently in an artistic versus nonartistic context. This approach was inspired by prior research that suggests that algorithmic aversion is larger in hedonic versus utilitarian domains (Longoni & Cian, 2022). While consumption of creative writing may generally be considered as hedonic, defined by the authors in the aforementioned study as "primarily affectively driven, based on sensory and experiential pleasure," we wondered whether manipulating the evaluation context to be more objective or utilitarian, defined in the aforementioned study as "cognitively driven, based on functional and instrumental goals," could mitigate the AI disclosure effects. We note that, while inspired by Longoni and Cian (2022), our approach is distinct in that we focus on the context under which evaluation takes place rather than changing the actual realm in which humans are interacting with the AI.

Procedure

We hypothesized that the AI disclosure penalty will be larger for writing samples in artistic contexts. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 150830, https://aspredicted.org/9T9_1MH). This experiment initially included 3,840 participants in the United States recruited via the Prolific platform, who were each compensated $1.25. After dropping participants who failed either of the two attention checks, the final sample included 3,590 participants (1,566 reported their gender as male, 1,961 as female, 46 as nonbinary or third gender, and two as other, and 15 declined to report; age = 18–79, $M_{\text{age}} = 34.7$, $SD_{\text{age}} = 12.5$).

The framework of this study was similar to that of Studies 1a–1e. We generated three prose writing samples using ChatGPT to serve as the study stimuli. In the study, all participants were asked to read and evaluate a writing sample and were randomly assigned participants to either be informed that the writing sample was written by "the AI tool ChatGPT" or by "a participant in a prior study." Across conditions, participants were randomly assigned to one of the three creative writing samples. Further, participants were randomly assigned to an artistic evaluation context, where they were asked to evaluate the writing sample based on its artistry, versus an objective evaluation context, where they were asked to evaluate the writing sample based on its objective quality and coherence.

After reading the writing samples, participants were asked to evaluate the writing sample. Rather than relying upon participant evaluations of creativity, quality, and enjoyment, we captured their evaluation of the writing sample using Berg's (2016) measure of audience appeal, as we felt this three-item measure was more likely to capture consumers' overall appreciation of the creative work. Participants rated the three items ("I [liked, appreciated, enjoyed] this writing sample") on a 7-point Likert-type scale (1 = *strongly disagree* to 7 = *strongly agree*). We collected the same demographic measures as in Studies 1a–1e. We also collected several measures to test as potential mechanisms and for exploratory analyses: We measured "eeriness" by adapting a measure from Ho and MacDorman's (2017) study, we collected measures of perceived understanding of human experiences and feelings from Liu and Sundar's (2018) study to capture the extent to which participants felt that AI had the psychological standing to be telling human stories, we collected the "heart" measures from Weisman et al.'s (2017) study to explore mechanisms related to humanness, and we collected perceived liking of the author to explore mechanisms related to familiarity.

Results

As preregistered, we explored the effect of AI disclosure using OLS regression. Our model used appeal as a dependent variable. As independent variables, we included AI disclosure, the topic, the artistic context, and the interaction of AI disclosure and the artistic context. We find a significant and negative AI disclosure effect ($p <$ .001) and do not find a significant interaction between AI disclosure and the context of the evaluation ($p = .573$).

Studies 3a–3e: How Does Perception of AI Shape AI Disclosure Effects?

Having found that the evaluation context did not appear to moderate the AI disclosure penalty, in the next series of studies, we considered how shaping participants' perceptions of the technology might influence AI disclosure effects. In Studies 3a and 3b, we consider whether manipulating beliefs regarding AI's capabilities may moderate the AI disclosure effect. This pair of studies was motivated by work that suggests that algorithmic literacy can mitigate algorithmic aversion by providing individuals with more context or information regarding how the algorithm works (see Burton et al., 2020, for a review). While work on algorithmic literacy often focuses on informing participants how best to use AI tools, we adapt this approach to suit the context of evaluating AI-generated creative writing. In particular, we focus on manipulating participants' beliefs regarding AI's capabilities to test how such beliefs may moderate AI disclosure effects. Then, in Studies 3c–3e, we tested whether humanizing the AI may moderate the AI disclosure penalty, building on research that suggests that anthropomorphizing technology can alter consumers' comfort with using it (e.g., Diel et al., 2021; Schanke et al., 2021).

The overarching framework of the studies is similar to Studies 1 and 2. Studies 3a–3e utilize AI-generated prose created using ChatGPT as stimuli. Participants were informed that they would be taking part in a study to examine how people evaluate written content. Across studies, all participants were asked to read and evaluate a writing sample and were randomly assigned participants to either be informed that the writing sample was written by an AI tool or by a participant in a prior study. After reading the writing samples, participants were asked to evaluate the appeal of the writing sample using the same measure as Study 2 from Berg (2016). However, in this subset of studies (3a–3e), we manipulate perceptions of the AI by providing participants with select information regarding the AI tool.

We collected the same demographic measures, measures of engagement, and potential mechanism measures as in Study 2 in Studies 3a–3c, though we did not capture eeriness in Study 3a, as the manipulation was less related to "uncanny valley" humanness effects. For Studies 3a–3c, we also collected participants' perceptions of AI's cognitive and emotional capabilities. In Studies 3d and 3e, we focused on two other potential mechanisms—perceived authenticity of the creative writing sample using a scale measure adapted from Park et al.'s (2016) study and perceived identity threat using a scale measure adapted from George et al.'s (2023) study. Below we provide more details on the individual studies.

Study 3a

Procedure. In this study, we considered whether manipulating beliefs regarding AI's capabilities would moderate the AI disclosure effect. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 155539, https://aspredicted.org/SPR_9V7). This experiment initially included 1,760 participants in the United States recruited via Prolific, who were each compensated $1.75. In this study, due to an error in the Qualtrics survey, 58 participants in the "pure control" condition were inadvertently assigned to review the same writing sample twice. Because we are unable to ascertain whether the evaluation recorded in Qualtrics is the initial evaluation (i.e., the one we would want to use), these participants are excluded from the study. We note that this is a deviation from the preregistration necessitated by the Qualtrics error. After dropping participants who failed either of two attention checks, the final sample included 1,526 participants (721 reported their gender as male, 773 as female, 21 as nonbinary or third gender, and seven as other, and four declined to report; age = 18–75, $M_{\text{age}} = 33.9$, $SD_{\text{age}} = 11.4$).

We generated three writing samples using ChatGPT that were used in the study. In addition to manipulating AI disclosure, we manipulated participants' perceptions of AI's capabilities by assigning them to one of four conditions: (a) an AI emotion condition wherein they are asked to first read and evaluate an article about the impressive emotional capabilities of AI technologies, (b) an AI cognition condition wherein they are asked to first read and evaluate an article describing the impressive cognitive capabilities of AI technologies, (c) a quantum computing condition wherein they are asked to first read and evaluate an article describing the impressive computing capabilities of quantum computing technologies, or (d) a pure control condition wherein they first evaluate the creative writing sample before being randomly assigned one of the other control articles to read and evaluate. These conditions were pretested to ensure that the AI emotion condition increases perceptions of AI's emotional capabilities relative to all other conditions. With the manipulation, we found that the AI emotion condition was successful in increasing the perception of AI's emotional capabilities relative to all other conditions ($p = .019$ for the test relative to the AI cognition condition, $p = .008$ relative to the pure control condition, and $p = .004$ relative to the quantum computing condition). Across disclosure conditions, participants were randomly assigned to one of the three creative writing samples.

Results. As preregistered, we explored the effect of AI disclosure using OLS regression. Our model used appeal as a dependent variable. As independent variables, we included AI disclosure, topic, information condition, and the interaction of AI disclosure and the perception manipulation condition. We find a significant and negative AI disclosure effect ($p = .001$) and do not find a significant interaction between AI disclosure and any of the perception conditions ($p = .816$, .267, and .581 for the AI cognition, pure control, and quantum computing conditions relative to the AI emotion condition).

Study 3b

Procedure. Having found that perceptions of AI's capabilities do not appear to moderate the AI disclosure penalty, we considered if information regarding the AI tool, rather than beliefs about its capabilities, may moderate the AI disclosure effect. In this study, we also include a manipulation that tests whether humanizing the AI tool may moderate the AI disclosure effect. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 159145, https://aspredicted.org/2WB_JJL). This experiment initially included 1,920 participants in the United States recruited via Prolific, who were each compensated $1.25. After dropping participants who failed either of the two attention check questions, the final sample included 1,799 participants (827 reported their gender as male, 937 as female, 29 as nonbinary or third gender, and one as other, and five declined to report; age = 18–79, $M_{\text{age}} = 33.3$, $SD_{\text{age}} = 12.0$).

We generated three writing samples using ChatGPT that were used in the study. We manipulated AI disclosure by informing participants that the author of the sample was either an AI tool or a past study participant. In addition, we manipulated information about the author or how the sample was written by assigning participants to one of four conditions: (a) a condition with no information about the author other than their identity, (b) a condition with objective information about the author, describing either the AI or the participant in descriptive terms, (c) a condition with humanizing information about the author, describing the AI or the participant in a humanizing manner (e.g., with a name and gender), or (d) a condition that describes AI as a tool and states either that the sample was written by an AI tool (in the AI author condition) or that the participant used AI as a tool in generating the sample (in the human author condition). In Condition (d), we note that participants were told that a human used AI to write the sample in both the AI and human conditions. As such, this condition can be viewed as a human–AI collaboration condition in which the primary (vs. secondary) author is being manipulated, unlike the other conditions wherein participants were told the author was either AI or human. The baseline reference was Condition (a), where no information was provided about the author other than their identity as AI or human. Across conditions, participants were randomly assigned to one of the three creative writing samples.

Results. As preregistered, we explored the effect of AI disclosure using OLS regression. Our model used appeal as a dependent variable. As independent variables, we included AI disclosure, topic, information provided, and the interaction of AI disclosure and the information condition. We find a significant and negative AI disclosure effect ($p = .019$). We do not find significant moderation in the objective or humanizing information conditions ($p = .653$ and .821, respectively). We find that the AI disclosure condition is positively and significantly moderated in the "tool" condition ($p = .026$); however, the tool condition also has a large and negative baseline effect ($p < .001$). Combining these two effects in a linear combination does not yield a significantly different result than the baseline AI disclosure effect. Effectively, these results suggest that in the context of human–AI collaboration, there is no significant difference in framing the human or AI tool as primary (vs. secondary) author—either framing yields a significant AI disclosure penalty compared with when participants think the sample is written by a human without AI assistance.

Study 3c

Procedure. In this study, we further consider whether humanizing the AI may moderate the AI disclosure penalty, using a stronger manipulation than we do in Study 3b. We hypothesized that the negative effect of AI disclosure would be larger when an AI tool is humanized, with the thought that anthropomorphizing an AI tool could generate a "backlash" effect as AI enters domains traditionally thought to be human (e.g., Diel et al., 2021). This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 161943, https://aspredicted.org/L23_R63). This experiment initially included 960 participants in the United States recruited via Prolific, who were each compensated $1.25. After dropping participants who failed either of the two attention check questions, the final sample included 864 participants (415 reported their gender as male, 434 as female, nine as nonbinary or third gender, and one as other, and five declined to report; age = 18–76, M_{age} = 33.1, SD_{age} = 11.3).

We generated three writing samples using ChatGPT that were used in the study. In addition to manipulating AI disclosure, we manipulated whether the author of the writing sample is humanized by randomly assigning participants to one of two conditions: (a) a condition with no information about the author other than their identity or (b) a condition with humanizing information about the author. Across conditions, participants were randomly assigned to one of the three creative writing samples. In addition to the measure of appeal used in earlier studies, we collected participant evaluations of the worth of the writing sample following Bellaiche et al. (2023) and a behavioral measure of willingness to purchase by capturing whether participants were willing to forgo a $0.10 bonus to read an additional writing sample.

Results. As preregistered, we explored the effect of AI disclosure on appeal, worth, and willingness to pay using OLS regression. As independent variables, we included AI disclosure, topic, humanizing information, and the interaction of AI disclosure and humanizing information. Considering appeal as a dependent variable, we do not find a significant baseline effect of AI disclosure (p = .264), but we do find that the interaction between AI disclosure and the humanizing information condition is negative and significant (p = .006), indicating that humanizing the AI increases the size of the AI disclosure penalty. Considering worth as a dependent variable, we find a marginal and negative effect of AI disclosure (p = .071) and no significant interaction between AI disclosure and humanizing information. Considering the behavioral measure of willingness to pay, we find neither a significant baseline effect of AI disclosure nor a significant interaction effect between AI disclosure and humanizing information.

Study 3d

Procedure. Having found preliminary evidence that humanizing the AI may exacerbate the AI disclosure penalty for creative writing, in Study 3d we run a constructive replication (including potential mediators of the backlash effect, the analyses for which are reported in the online Supplemental Appendix). We again hypothesized that the negative effect of AI disclosure would be larger when an AI tool is humanized. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 164689, https://aspredicted.org/6D9_SF5). This experiment initially included 960 participants in the United States recruited via Prolific, who were each compensated $1.40. After dropping participants who failed either of the two attention check questions, the final sample included 909 participants (436 reported their gender as male, 455 as female, and 16 as nonbinary or third gender, and two declined to report; age = 18–99, M_{age} = 34.4, SD_{age} = 11.9). We generated three writing samples using ChatGPT that were used in the study. In addition to manipulating AI disclosure, we manipulated whether the author of the writing sample is humanized by randomly assigning participants to the same two conditions as Study 3c. Across conditions, participants were randomly assigned to one of the three creative writing samples.

Results. As preregistered, we explored the effect of AI disclosure on appeal using OLS regression. As independent variables, we included AI disclosure, topic, humanizing information, and the interaction of AI disclosure and humanizing information. In contrast to the prior study, we find a negative and significant baseline effect of AI disclosure (p = .002) but do not find a significant negative interaction between AI disclosure and humanizing information (p = .529).

Study 3e

Procedure. Given the disconnect between the results of Studies 3c and 3d, we ran a constructive replication to identify whether humanizing an AI may exacerbate the AI disclosure penalty for creative writing. We again hypothesized that the negative effect of AI disclosure would be larger when an AI tool is humanized and used a stronger manipulation of humanization, adding more language to humanize the AI tool. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 166826, https://aspredicted.org/VNX_4JG). This experiment initially included 960 participants in the United States recruited via the Prolific platform, who were each compensated $1.40. After dropping participants who failed either of the two attention checks, the final sample included 908 participants (378 reported their gender as male, 519 as female, three as nonbinary or third gender, and one as other, and three declined to report; age = 18–78, M_{age} = 33.6, SD_{age} = 11.6).

We generated three writing samples using ChatGPT that were used in the study. In addition to manipulating AI disclosure, we manipulated whether the author of the writing sample is humanized by randomly assigning participants to one of two conditions: (a) a condition with no information about the author other than their identity or (b) a condition with humanizing information about the author. Relative to the prior study, we used even more human language to describe the AI and to further increase perceived humanization of the AI. Across conditions, participants were randomly assigned to one of the three creative writing samples.

Results. As preregistered, we explored the effect of AI disclosure on appeal using OLS regression. As independent variables, we included AI disclosure, topic, humanizing information, and the interaction of AI disclosure and the humanizing information. In contrast to Study 3c but consistent with Study 3d, we find a significant baseline effect of AI disclosure (p = .015) but do not find a significant negative interaction between AI disclosure and humanizing information (p = .922).

Studies 4a and 4b: How Does AI Disclosure Affect Ambivalence About the Content?

Although results from Study 3c suggest that humanizing an AI exacerbates the AI disclosure penalty, Studies 3d and 3e showed that these results are not robust to replication. We thus again took another tack and, motivated by the inconsistency of our prior findings, sought to understand whether AI disclosure may affect ambivalence, defined as the presence of both positive and negative evaluations of the same object (e.g., Thompson et al., 1995). In Study 4a, we sought to identify whether there is a baseline relationship between AI disclosure and ambivalence, before testing to see whether humanizing an AI moderates the relationship between AI disclosure and ambivalence in Study 4b.

The framework of the studies is similar to our prior studies, utilizing AI-generated prose as stimuli and informing participants that they will be taking place in a study to examine how people evaluate written content. In both studies, participants were asked to read and evaluate a writing sample and were randomly assigned to be informed that the writing sample was written either by an AI tool or by a participant in a prior study. After reading the writing sample, participants were asked to evaluate the appeal of the writing sample using the same (Berg, 2016) measure as in earlier studies. Further, in these studies, we collect a measure of ambivalence. Following Thompson et al. (1995), we measure ambivalence by capturing both positive and negative evaluations of a writing sample using the prompt "Considering only the (positive/negative) qualities of the writing sample and ignoring its (negative/positive) ones, evaluate how (positive/negative) its (positive/negative) qualities are on the following 4-point scale." We then calculated ambivalence using the Griffin formula Ambivalence = $(P + N)/2 - \text{abs}(P - N)$. We collected the same demographic data, measures of engagement, mechanism measures, and a broad range of exploratory measures as in Study 3b.

Study 4a

Procedure. We examined whether there is a baseline relationship between AI disclosure and ambivalence, hypothesizing that the AI disclosure would increase feelings of ambivalence. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 169423, https://aspredicted.org/ZKR_K9P). This experiment included 480 participants in the United States recruited via Prolific, who were each compensated $1.25. After dropping participants who failed either of the two attention checks, the final sample included 423 participants (169 reported their gender as male, 243 as female, and nine as nonbinary or third gender, and two declined to report, age = 18–82, $M_{\text{age}} = 31.8$, $SD_{\text{age}} = 10.9$). We generated three writing samples using ChatGPT that were used in the study. We utilize a simple AI disclosure condition, and across conditions, participants were randomly assigned to one of the three creative writing samples.

Results. As preregistered, we explored the effect of AI disclosure on ambivalence using OLS regression. As independent variables, we included AI disclosure and the topic. We do not find a significant relationship between ambivalence and AI disclosure (p = .401). We do, however, continue to find a significant and negative relationship between AI disclosure and appeal ($p < .001$).

Study 4b

Procedure. We aimed to test whether humanizing the AI moderates the relationship between AI disclosure and ambivalence in Study 4b. We hypothesized that humanizing an AI would positively moderate the relationship between AI disclosure and ambivalence, as we felt it might generate a backlash effect while simultaneously making the AI feel more relatable or human. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 170306, https://aspredicted.org/3VG_NPD). This experiment initially included 1,440 participants in the United States recruited via Prolific, who were each compensated $1.25. After dropping participants who failed either of the two attention checks, the final sample included 1,280 participants (556 reported their gender as male, 696 as female, 19 as nonbinary or third gender, and three as other, and six declined to report; age = 18–89, $M_{\text{age}} = 33.7$, $SD_{\text{age}} = 11.5$). This study used the same three writing samples as Study 4a. In addition to the AI disclosure manipulation, we randomly assigned participants to a condition wherein the AI was humanized (as in Study 3e). Across conditions, participants were randomly assigned to one of the three creative writing samples.

Results. As preregistered, we explored the effect of AI disclosure on ambivalence. As independent variables, we included AI disclosure, topic, humanizing information, and the interaction between AI disclosure and humanizing information. We do not find a significant relationship between ambivalence and AI disclosure (p = .771) or the interaction of AI disclosure and humanizing information (p = .924). Considering appeal as the dependent variable (rather than ambivalence), results are consistent with Studies 3d, 3e, and 4a: We find a significant and negative baseline relationship of AI disclosure ($p < .001$) but do not find a significant interaction between AI disclosure and humanizing information (p = .436).

Study 5a–5c: How Does a Human-in-the-Loop Framing Shape AI Disclosure Effects?

Having found no evidence that humanizing AI leads to moderation of the AI disclosure effect (for either enjoyment/appeal or ambivalence), we again pivoted to try and find another intervention that may moderate the relationship between AI disclosure and evaluations. We returned to the significant moderation in Study 3b. In Study 3b, we found that AI disclosure was positively and significantly moderated by describing the AI as a tool used by humans, but the baseline tool condition also had a large and negative baseline effect. In combination, this study suggested that there was no significant difference between describing a sample as human written with the assistance of AI and describing a sample as AI written with the AI framed as a tool used by humans. We decided to probe this further by testing whether other ways of bringing humans in the loop may moderate the AI disclosure penalty (e.g., Dietvorst et al., 2018; Horton et al., 2023; Mellamphy, 2021).

As with prior studies, participants were informed that they would be taking place in a study to examine how people evaluate written content. We then had participants read and evaluate writing samples, manipulating whether they were informed that the sample was written by an AI tool, a participant in a prior study, or a participant using an AI tool. As stimuli, we use the same three ChatGPT-generated writing samples from Studies 4a and 4b in Studies 5a and 5b and human-written prose in Study 5c. We collected the same measure of appeal, demographics, measures of engagement, mechanism measures, and a broad range of exploratory measures as in Study 4b, except we did not collect measures of humanness or perceptions of whether the author understands human feelings or experiences in Studies 5b and 5c. We also captured measures of perceived authenticity adapted from Park et al.'s (2016) study as in Study 3e across Studies 5a–5c.

Study 5a

Procedure. We examined whether describing the AI as a tool used by humans could moderate the AI disclosure penalty. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 172429, https://aspredicted.org/Y9J_7BN). This experiment initially included 960 participants in the United States recruited via Prolific, who were each compensated $1.25. After dropping participants who failed either of the two attention checks, the final sample included 856 participants (358 reported their gender as male, 480 as female, 14 as nonbinary or third gender, and two as other, and two declined to report; age = 18–78, $M_{\text{age}} = 33.5$, $SD_{\text{age}} = 11.7$).

We manipulated AI disclosure by randomly assigning participants to either (a) a condition wherein the participant is told that the writing sample was written by an AI model or (b) a condition wherein the participant is told that the writing sample was written by a participant in a prior study. We manipulated the tool conditions by randomly assigning participants to either (a) a condition that describes AI as a tool and states that the sample was written by an AI tool or that a participant used AI as a tool in generating the sample or (b) a baseline condition that included no such information. Across conditions, participants were randomly assigned to one of the three creative writing samples.

Results. As preregistered, we explored the effect of AI disclosure on appeal using OLS regression. As independent variables, we included AI disclosure, topic, tool condition, and the interaction between AI disclosure and tool condition. We continue to find a significant and negative baseline relationship of AI disclosure (p = .001) but do not find a significant interaction between AI disclosure and tool condition (p = .114). We do find a significant and negative baseline effect of tool condition (p = .003). As in Study 3b, it does not appear that describing the writing sample as written by an AI tool used by a human versus as written by a human using AI as a tool results in a significant difference in appeal.

Study 5b

Procedure. In Study 5b, we build upon Study 5a to consider how describing a creative writing sample as written through AI–human collaboration versus written by an AI versus written by a human affects audience evaluations of appeal. This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 173344, https://aspredicted.org/5B6_CZF). This experiment initially included 720 participants in the United States recruited via Prolific, who were each compensated $1.25. After dropping participants who failed either of two attention checks, the final sample included 665 participants (281 reported their gender as male, 369 as female, and 12 as nonbinary or third gender, and three declined to report; age = 18–77, $M_{\text{age}} = 33.9$, $SD_{\text{age}} = 12.2$). We randomly assigned each participant to one of three disclosure conditions: (a) a condition wherein they are told that the writing sample was generated by an AI model, (b) a condition wherein they are told that the writing sample was written by a participant in a prior study, and (c) a condition wherein they are told that the writing sample was written by a participant in a prior study using an AI model. Across conditions, participants were randomly assigned to one of the three writing samples.

Results. As preregistered, we explored the effect of AI disclosure on appeal using OLS regression. As independent variables, we included the disclosure condition and the topic of the writing sample. We continue to find a significant and negative baseline effect of AI disclosure (p = .008), and we also find a significant and negative baseline effect of collaboration disclosure (p < .001). While the magnitude of the effect estimate is larger for the collaboration disclosure condition, the difference between the two estimates is not significant (p = .359). Accordingly, both collaboration disclosure and AI disclosure have a similar negative effect on evaluations of appeal.

Study 5c

Procedure. Thus far, our studies establishing the effect of AI disclosure rely upon AI-generated writing samples. One concern could be that the quality of such samples is systematically worse or different from those written solely by humans, and such quality differences could importantly shape the results we document. To investigate whether this is the case, we replicate Study 5b using award-winning human-written prose (five winning short stories from the Writers' Digest Short Stories competition from 2018 through 2022). This experiment was preregistered on https://AsPredicted.org (AsPredicted No. 178258, https://aspredicted.org/YQ2_F2Q). The experiment initially included 1800 participants in the United States recruited via the Prolific platform, who were each compensated $1.50. After dropping participants who failed either of the two attention checks, the final sample included 1,572 participants (646 reported their gender as male, 897 as female, 22 as nonbinary or third gender, and one as other, and six declined to report; age = 18–80, $M_{\text{age}} = 33.5$, $SD_{\text{age}} = 11.7$). We utilized the same disclosure conditions as in Study 5b. Across conditions, participants were randomly assigned to one of five creative writing samples.

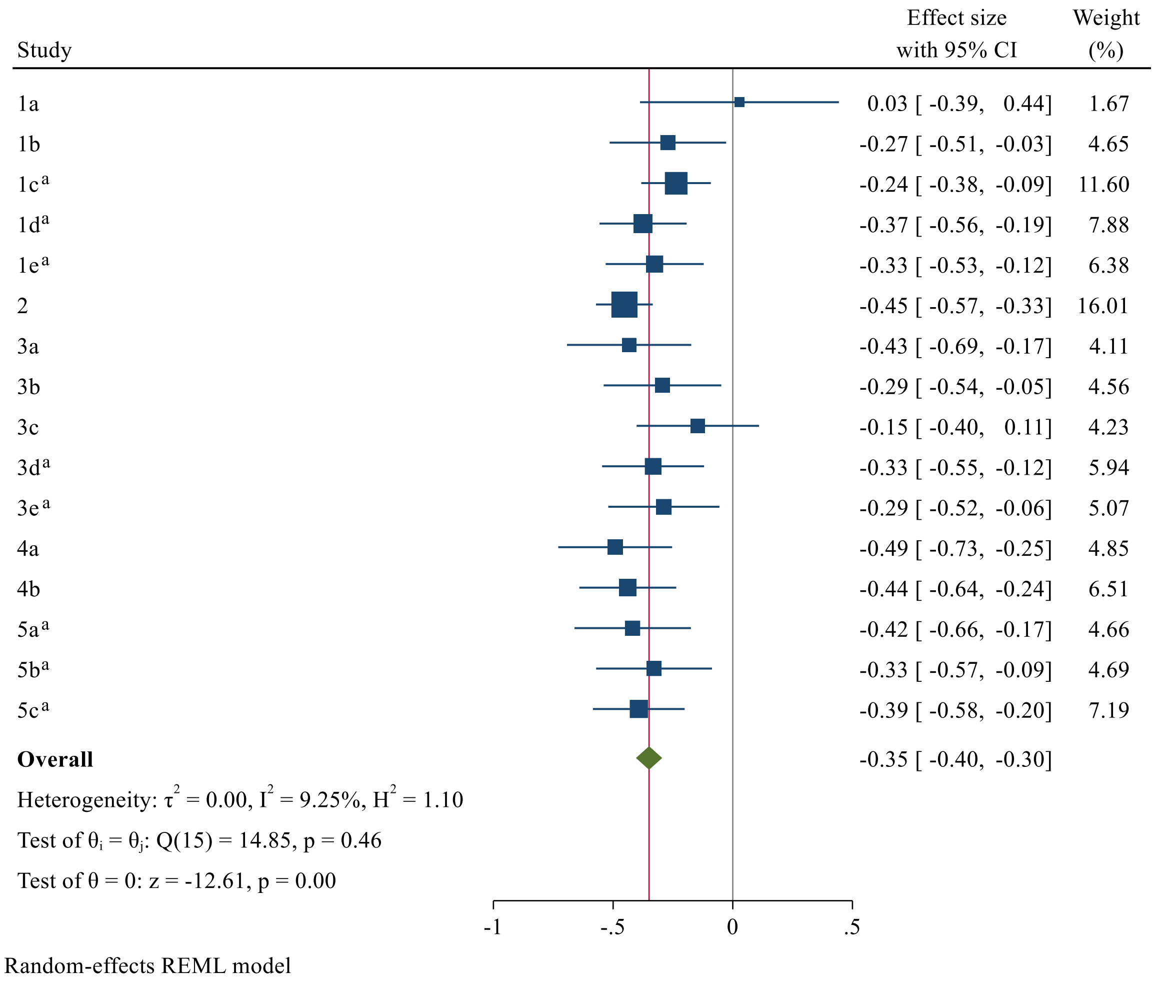
Results. As preregistered, we explored the effect of AI disclosure on appeal using OLS regression. As independent variables, we included the disclosure condition and the topic of the writing sample. We continue to find a significant and negative baseline effect of AI disclosure (p < .001) and collaboration disclosure (p < .001). The estimates of the effects of these two conditions are nearly identical in magnitude (−0.393 vs. −0.390 for the AI and collaboration disclosure conditions, respectively), and the difference between the two estimates is not significant (p = .976).

Meta-Analysis of the AI Disclosure Penalty Across Studies

Across our 16 studies (N = 27,491), we found consistent evidence that evaluations of creative writing samples were negatively affected by disclosing the use of AI in the creation of these samples. To assess the baseline AI disclosure penalty across studies, we conducted a meta-analysis. Figure 1 presents the baseline effect of AI disclosure on participant evaluations for each study, as well the results of this meta-analysis. For the meta-analysis, we estimate the effect of AI disclosure in each study using an ordinary least squares linear regression. Studies 1a–1e ask participants to evaluate the quality, creativity, and enjoyment of the writing sample using a 1–7 scale, while Studies 2–5c utilize a measure of audience appeal (Berg, 2016). For ease of interpretation, in comparing results across studies in the meta-analysis, we focus on evaluations of enjoyment in Studies 1a–1e, as enjoyment is more closely linked to audience appeal than creativity or quality (results are qualitatively similar using evaluations of creativity or quality). Aggregating across studies, we find a statistically strong and negative relationship between AI disclosure and participant evaluations ($p < .001$). In 14 of 16 studies, the effect of AI disclosure is negative and significant at $p < .05$. The two exceptions are Studies 1a and 3c. In both these studies, using a specification that does not include interactions or

Figure 1

Forest Plot of Meta-Analysis Results



Note. This figure presents a forest plot obtained by conducting a meta-analysis to summarize the average effect of artificial intelligence disclosure for the studies presented in Table 1. Studies 1–5 use participant enjoyment as a dependent variable; Studies 6–16 use appeal as measured by Berg (2016). Results are robust, conducting subgroup analyses by dependent variable. Estimated effects presented are the baseline artificial intelligence disclosure effects obtained using ordinary least squares regression models as specified in the preregistration for each study, except for Study 1a, which preregistered analysis of variance rather than OLS (the baseline penalty was significant with analysis of variance, but we use OLS here for consistency with the other studies). The meta-analysis is conducted using STATA's meta command. We use the default random-effects specification, which assumes that collected studies represent a random sample from a larger population of studies. The red line represents the meta-analysis effect estimates, and the gray line represents a null effect. Point estimates and confidence intervals for each study are presented with the side of each marker representing the relative weight of the study in constructing the overall meta-analysis effect estimate (presented with the green diamond marker). The width of the green diamond marker indicates the confidence interval of the overall meta-analysis effect estimate. CI = confidence interval; REML = restricted maximum likelihood. See the online article for the color version of this figure.

[a] Study wherein we collect data on authenticity and test for mediation of the artificial intelligence disclosure effect using authenticity.

using ANOVA with the preregistered specification, we do find the baseline negative effect of AI disclosure seen in the other studies. The result of Cochran's *Q* test, which produces a *p* value of .46, indicates that the variability in effect sizes across experiments is not significantly greater than what would be expected by chance alone. This suggests that the observed AI disclosure penalty is consistent and reliable across studies.

We note that the large sample and the number of studies across March 2023 through June 2024, a period featuring large advances in generative AI that could generate creative writing, allow us to evaluate whether and how AI disclosure effects may have shifted as this technology continued to enter the public eye and perception evolved regarding its capabilities and potential uses (Walt, 2023). The consistency of estimates across this period indicates that such changes did not meaningfully shape AI disclosure effects with respect to creative writing. Further, we note that the size of the AI disclosure effect, while small, is not negligible in magnitude. Across studies, AI disclosure decreased evaluations by 6.2% on average, with effect sizes calculated as the unstandardized coefficient estimate for AI disclosure divided by the sample mean for the dependent variable in each study. Put differently, the average Cohen's *d*, constructed using simple *t* tests to measure the effect of AI disclosure within each study, across studies is 0.24.[2]

Mediation Analyses: Perceived Authenticity as a Key Driver of the AI Disclosure Penalty

Across experiments, we collected and tested a variety of potential mediators related to perceptions of the written content, such as perceived authenticity, humanness, emotion, profundity, meaningfulness, worth, effort, liking and familiarity, and eeriness, as well as related to perceptions of the AI/author, such as perceived understanding of human feelings, perceived understanding of human experiences, perceived cognitive capabilities, and perceived emotional capabilities. The measures we explored as mediators were often motivated by the particular tests in the studies and thus were not always consistent across studies. The overarching logic underlying these mediators was centered around choosing items linked to the human element of creative content and metrics that prior research on AI or automation suggested may be shaped by the use of AI.

We tested whether the relationship between AI disclosure and evaluation is mediated by each potential mechanism measure considered in each study, using Hayes's (2013) bootstrapping method to estimate the indirect effects—Table 2 contains these estimates and corresponding 95% confidence intervals. Overall, we find that many of the mediators considered appear to account for the negative relationship between AI disclosure and participant evaluations, which makes sense given that these potential mediators were selected based on prior related research.

Although we find significant evidence for several different mediators, from a conceptual standpoint, we propose that perceived authenticity offers the most elegant explanation for why the AI disclosure penalty happens across a wide array of content and contexts, particularly because authenticity seems to subsume many of the other significant mediators in our studies and related studies by other scholars (e.g., Bellaiche et al., 2023), such as profundity, meaningfulness, effort, and the AI/author's understanding of human feelings and experiences. From an empirical standpoint, perceived authenticity was a consistently strong mediator of the AI disclosure penalty in every study in which we measured it. That is, in all eight studies in which we collected data on perceived authenticity, we found that AI disclosure was negatively related to perceived authenticity, the indirect effect of perceived authenticity was significant, and controlling for perceived authenticity rendered the negative effect of AI disclosure insignificant. Further, we find that the positive relationship between perceived authenticity and evaluations in our mediation models persists even when controlling for all other potential mediators for which we collected data.

To assess the overall strength of evidence for perceived authenticity as a mediator of the AI disclosure penalty, we conducted a meta-analysis that summarizes the total, direct, and indirect effects of AI disclosure on evaluations, mediated by perceived authenticity, across all the studies in which perceived authenticity was measured. In Table 3, we present the meta-analysis results, including a breakdown of results for each study that considered authenticity as a potential mediator. The meta-analysis reveals that the indirect effect of AI disclosure on evaluations through perceived authenticity is statistically significant (indirect effect = −0.292, $p < .001$). Further, the direct effect of AI disclosure on evaluations, controlling for perceived authenticity, is no longer significant and close to zero (direct effect = −0.034, $p = .240$). Figure 2 depicts the relationship between AI disclosure, perceived authenticity, and evaluations.

General Discussion

Across 16 preregistered experiments, we document evidence of a robust AI disclosure penalty. Participant evaluations of creative writing samples decrease when they believe that the writing samples were written by, or with the help of, an AI model rather than a human author without the use of AI. This AI disclosure penalty is mediated by perceived authenticity, suggesting that, at least at the time of our study, people tend to view AI-generated creative goods as inauthentic and therefore less worthy of their appreciation. This effect is remarkably persistent, holding across the time period covered in our study, across different evaluation metrics, contexts, kinds of written content, and across interventions derived from prior research aimed at moderating the penalty. Specifically, we designed and tested a variety of interventions with the intention to moderate the AI disclosure penalty, building on work showing that reactions to AI can differ in hedonic versus utilitarian domains (Longoni & Cian, 2022), when participants' perceptions of an AI's capabilities are higher versus lower (Bellaiche et al., 2023), when an AI tool is more versus less humanized (e.g., Burton et al., 2020; Schanke et al., 2021), and when a human is "in the loop" (e.g., Dietvorst et al., 2018; Hong et al., 2022; Horton et al., 2023). Despite our deliberate attempts, we were unable to find consistent moderation of the effect, suggesting that the AI disclosure penalty with respect to creative writing is, at least at this time, quite robust.

Through this research, we advance a small but growing body of literature that integrates work on the effects of AI disclosure (e.g., Bellaiche et al., 2023; Horton et al., 2023; Tigre Moura et al., 2023)

[2] In calculating Cohen's *d*, we exclude observations assigned to the human–AI collaboration disclosure in Studies 5b and 5c to construct a clean comparison between AI disclosure and blind conditions.

Table 2

Summarizing Indirect Effect Estimates of Mediators Across Studies

Study	Authenticity	Humanness	Profundity	Emotion	Story	Meaningfulness	Effort	Worth
1c	−0.241*** [−0.328, −0.154]	−0.222*** [−0.305, −0.139]	−0.159*** [−0.250, −0.069]	−0.214*** [−0.316, −0.111]	−0.126** [−0.215, −0.037]	−0.191*** [−0.288, −0.095]	−0.540*** [−0.618, −0.462]	−0.132** [−0.218, −0.047]
1d	−0.287*** [−0.403, −0.171]	−0.202*** [−0.296, −0.108]	−0.211*** [−0.322, −0.101]	−0.330*** [−0.453, −0.207]	−0.134* [−0.253, −0.016]	−0.259*** [−0.378, −0.139]	−0.557*** [−0.652, −0.461]	−0.139* [−0.250, −0.029]
1e	−0.211*** [−0.332, −0.090]	−0.118** [−0.203, −0.033]	−0.189** [−0.312, −0.066]	−0.202** [−0.338, −0.065]	−0.159* [−0.294, −0.024]	−0.231** [−0.370, −0.092]	−0.609*** [−0.706, −0.511]	−0.223*** [−0.348, −0.098]

Study	Authenticity	Humanness	Eeriness	Author's liking	Author's understanding of feeling	Author's understanding of emotion	Author's cognitive capability	Author's emotional capability	Identity threat
2		−0.347*** [−0.400, −0.295]	−0.177*** [−0.217, −0.136]	−0.482*** [−0.574, −0.390]	−0.500*** [−0.571, −0.429]	−0.426*** [−0.495, −0.357]			
3a		−0.044 [−0.098, 0.011]		0.002 [−0.038, 0.043]	0.01 [−0.052, 0.071]	0.021 [−0.041, 0.083]	0.019 [−0.069, 0.106]	0.016 [−0.046, 0.079]	
3b		−0.331*** [−0.428, −0.233]	−0.105 [−0.209, 0.000]	−0.257** [−0.446, −0.068]	−0.508*** [−0.642, −0.375]	−0.459*** [−0.590, −0.328]	−0.438*** [−0.573, −0.303]	−0.573*** [−0.713, −0.433]	
3d	−0.337*** [−0.471, −0.203]								−0.001 [−0.008, 0.006]
3e	−0.300*** [−0.444, −0.156]								−0.001 [−0.012, 0.011]
4a		−0.241*** [−0.345, −0.138]	−0.281*** [−0.404, −0.158]	−0.298*** [−0.462, −0.134]	−0.453*** [−0.595, −0.311]	−0.448*** [−0.587, −0.308]	−0.367*** [−0.504, −0.230]	−0.467*** [−0.613, −0.322]	
4b		−0.364*** [−0.454, −0.274]	−0.215*** [−0.314, −0.116]	−0.429*** [−0.585, −0.274]	−0.507*** [−0.620, −0.394]	−0.460*** [−0.568, −0.351]	−0.531*** [−0.653, −0.409]	−0.608*** [−0.728, −0.488]	
5a	−0.401*** [−0.550, −0.253]	−0.336*** [−0.448, −0.225]	−0.234*** [−0.352, −0.116]	−0.429*** [−0.622, −0.235]	−0.441*** [−0.577, −0.305]	−0.360*** [−0.492, −0.229]	−0.398*** [−0.539, −0.257]	−0.488*** [−0.631, −0.345]	
5b	−0.341*** [−0.490, −0.192]		−0.185** [−0.305, −0.065]	−0.274** [−0.456, −0.091]			−0.364*** [−0.511, −0.218]	−0.479*** [−0.627, −0.332]	
5c	−0.322*** [−0.429, −0.215]		−0.109** [−0.178, −0.040]	−0.315*** [−0.469, −0.161]			−0.532*** [−0.652, −0.412]	−0.606*** [−0.725, −0.488]	

Note. In this table, we provide estimated indirect effects and 95% confidence intervals for all potential mediators considered in each study. We utilize enjoyment as the dependent variable in Studies 1c–1e and appeal as the dependent variable in Studies 2–5c. This table only includes studies wherein we tested for mediation. Mediator variables refer to participants' views of the writing sample unless "author" is in the label.

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 3

Meta-Analysis Summarizing the Total, Direct, and Indirect Effects of Artificial Intelligence Disclosure on Evaluations as Mediated by Authenticity

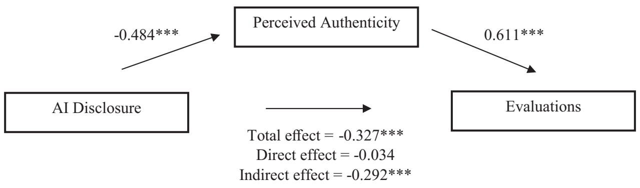
	Total effect				Direct effect				Indirect effect			
Study	Effect size	Lower bound	Upper bound	Weight	Effect size	Lower bound	Upper bound	Weight	Effect size	Lower bound	Upper bound	Weight
1c	−0.237	−0.382	−0.092	23.3%	0.004	−0.113	0.122	23.2%	−0.241	−0.328	−0.154	23.7%
1d	−0.375	−0.557	−0.193	14.9%	−0.088	−0.229	0.053	16.0%	−0.287	−0.403	−0.171	13.4%
1e	−0.326	−0.531	−0.121	11.7%	−0.115	−0.280	0.050	11.7%	−0.211	−0.332	−0.090	12.2%
3d	−0.333	−0.546	−0.120	10.8%	0.004	−0.166	0.174	11.1%	−0.337	−0.471	−0.203	10.0%
3e	−0.288	−0.520	−0.056	9.1%	0.012	−0.172	0.197	9.4%	−0.300	−0.444	−0.156	8.7%
5a	−0.418	−0.661	−0.175	8.3%	−0.017	−0.214	0.181	8.2%	−0.401	−0.550	−0.253	8.1%
5b	−0.329	−0.571	−0.087	8.4%	0.012	−0.184	0.208	8.3%	−0.341	−0.490	−0.192	8.1%
5c	−0.393	−0.584	−0.201	13.4%	−0.071	−0.233	0.092	12.1%	−0.322	−0.429	−0.215	15.8%
θ	**−0.327**	**−0.397**	**−0.257**	**100%**	**−0.034**	**−0.090**	**0.023**	**100%**	**−0.292**	**−0.335**	**−0.250**	**100%**
Test				*p*				*p*				*p*
Test of θ = 0				<.001				.240				<.001
Test of homogeneity				.899				.906				.510

Note. The summary low containing the averaged effect across studies appears in bold. This table includes all studies for which data were collected on participant evaluations of authenticity were collected. Lower and upper bounds refer to the 95% confidence interval for each estimate. Figure 2 graphically displays the mediation pathway tested.

with literature on the impact of AI on creative work (e.g., Amabile, 2020; Doshi & Hauser, 2024). The sticky AI disclosure penalty we document poses implications for the integration of AI in creative fields. While previous research has offered preliminary evidence that AI disclosure can have negative effects on evaluations of creative content, our research builds upon extant work by revealing the consistency and robustness of the AI disclosure penalty in the context of creative writing. Our results suggest a persistent bias against AI-generated and AI-assisted content that endures even through interventions that have mitigated similar biases in prior research. The stickiness of this effect raises questions regarding whether and to what extent aversive reactions to AI disclosure are similar to and different from other forms of algorithmic aversion. Further, our inability to find moderation using manipulations that appear to have mitigated AI disclosure effects in other settings (e.g., Bellaiche et al., 2023; Horton et al., 2023) raise questions regarding how such effects may differ across different forms of creative goods or may evolve as the technology advances and its place in our society changes. These results suggest that human reactions to the use of AI in the creation of creative writing may, at least at this point in time, trigger different psychological responses in individuals than the use of AI in other domains. As AI tools become increasingly prevalent in creative work, understanding this bias is crucial for helping stakeholders navigate the challenges that must be overcome to harness the full potential of human–AI collaboration. Further, our findings pose practical implications for creative producers using AI, which are especially pertinent as the U.S. Congress considers AI disclosure legislation (AI Disclosure Act of 2023, 2023; H.R. 3831, 118th Congress). If such legislation mandates the disclosure of AI involvement in creative work, it could reify negative biases toward AI-generated content, potentially affecting the reception of creative works and the livelihood of creators who could otherwise benefit from using AI tools.

Figure 2

Mediation of the Relationship Between AI Disclosure and Evaluations



Note. Standardized regression coefficients for the relationship between AI disclosure and appeal as mediated by participants' evaluations of authenticity obtained from the meta-analysis presented in Table 3. AI = artificial intelligence.

*** $p < .001$.

Our research also contributes to the growing literatures on the impact of AI (e.g., Amabile, 2020; Doshi, 2025; Tong et al., 2021) and evaluative biases (e.g., Berg, 2016; Lucas & Nordgren, 2020) in creative work. We build upon research documenting that AI disclosure results in lower evaluations of creative work (e.g., Bellaiche et al., 2023; Horton et al., 2023) by extending this finding to the context of creative writing, documenting that this relationship is mediated by perceived authenticity, and demonstrating that the AI disclosure penalty is persistent across different kinds of creative writing and surprisingly unresponsive to interventions that mitigate algorithmic aversion and AI disclosure effects in previous literature. More broadly, this work contributes to literature on social psychology by considering how perceivers may respond to a nonhuman, AI identity of a creator, distinct from prior work that largely explores dimensions of identity within the human category (e.g., Cuddy et al., 2008), and deepens our understanding of how individuals assess whether an AI possesses mental life (e.g., Gray et al., 2007).

It is important to document the limitations of this work. We have studied the effect of AI disclosure on evaluations in one specific domain (creative writing), and although we manipulated the content in several ways (e.g., poem vs. prose, first- vs. third-person perspective, many different topics), there may be other dimensions that could be relevant to the AI disclosure penalty, and it is possible that there are some forms of creative goods that may not elicit such strong aversive reactions. We are also careful to note that our study does not address whether and in what circumstances output created by an AI tool may be more or less creative than output created by a human but instead solely focuses on the effects of AI disclosure holding the content of the output constant. Finally, while our results are stable over the 15-month study period (March 2023 to June 2024), it is important to note that the AI disclosure effects we document may evolve over time, both as the sophistication of AI tools increases and as we become more accustomed and inured to the idea of AI-generated creative goods.

With that said, these findings contribute to existing literature on algorithmic aversion (e.g., Dietvorst et al., 2015; Logg et al., 2019) and to a rapidly growing literature on AI disclosure (e.g., Bellaiche et al., 2023; Horton et al., 2023; Jago, 2019; Tigre Moura et al., 2023). We build upon existing work that documents an AI disclosure effect on evaluations of visual art and music and document a similar effect on evaluations of creative writing. Further, we extend such work by conducting a broad exploration to try and identify heterogeneity in such an effect. Our inability to find consistent moderation of the AI disclosure penalty, even when utilizing interventions that have mitigated algorithmic aversion in previous literature, emphasizes the persistence of algorithmic aversion in the domain of creative writing and hints that similarly stubborn AI disclosure penalties may emerge in other creative domains as AI tools continue to proliferate. Further, the findings highlight a practical consideration for creators considering when and how to use and disclose the use of AI in the production of artistic products. Specifically, creators face a difficult trade-off: While AI may enhance the production of their creative work (Doshi & Hauser, 2024), disclosing its use may negatively impact audience perceptions, putting creators in a thorny bind as they navigate decisions about transparency and commercial success in a world where AI promises to become increasingly prevalent.

Constraints on Generality

The findings of this study should be interpreted with attention to their generalizability. Our experiments were conducted with participants recruited through the online platform Prolific, which tends to yield a sample that is more educated, tech savvy, and Western centric than the general population. As a result, it is unclear if the observed AI disclosure penalty will generalize to other populations with different levels of exposure to AI-generated content or varying cultural attitudes toward AI and creativity. Additionally, our study focuses on written creative content in English, limiting generalizability to other forms of creative output, such as visual art, music, or performance-based media, where perceptions of authenticity and human involvement may differ, or to non-English language settings or cultures. Furthermore, our results reflect attitudes toward AI during a specific time period (March 2023 to June 2024), a period of rapid advancements in AI capabilities and shifting societal perceptions of AI's role in creative work. While our results are persistent across this time period, it remains an open question (and a fruitful avenue for future study) whether the AI disclosure penalty will persist, diminish, or even reverse as AI-generated content becomes more ubiquitous and consumers become more accustomed to its presence in creative fields. Future research should explore whether these effects hold in non-English contexts, across diverse cultural backgrounds, and as AI technology and its societal reception evolve.

References

AI Disclosure Act of 2023, H.R. 3831, 118th Congress (2023). https://www.congress.gov/bill/118th-congress/house-bill/3831/text

Amabile, T. M. (2020). Creativity, artificial intelligence, and a world of surprises. *Academy of Management Discoveries*, *6*(3), 351–354. https://doi.org/10.5465/amd.2019.0075

Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N., Christensen, A., & Seli, P. (2023). Humans versus AI: Whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research: Principles and Implications*, *8*(1), Article 42. https://doi.org/10.1186/s41235-023-00499-6

Berg, J. M. (2016). Balancing on the creative highwire: Forecasting the success of novel ideas in organizations. *Administrative Science Quarterly*, *61*(3), 433–468. https://doi.org/10.1177/0001839216642211

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Chow, A. R. (2023, December 20). 4 ways AI transformed music, movies and art in 2023. *TIME*. https://time.com/6343945/ai-music-movies-art-2023/

Correia, S. (2023a). *FTOOLS: Stata module to provide alternatives to common Stata commands optimized for large datasets* [Computer software]. Statistical Software Components. https://ideas.repec.org//c/boc/bocode/s458213.html

Correia, S. (2023b). *REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects* [Computer software]. Statistical Software Components. https://ideas.repec.org//c/boc/bocode/s457874.html

Coscarelli, J. (2023, April 19). An A.I. hit of fake 'drake' and 'the weeknd' rattles the music world. *The New York Times*. https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, *40*, 61–149. https://doi.org/10.1016/S0065-2601(07)00002-0

Diel, A., Weigelt, S., & Macdorman, K. F. (2021). A meta-analysis of the uncanny valley's independent and dependent variables. *Journal of Human-Robot Interaction*, *11*(1), 1–33. https://doi.org/10.1145/3470742

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

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*(2), 192–203. https://doi.org/10.1037/1040-3590.18.2.192

Doshi, A. R. (2025). How high-performance outliers affect relative entrepreneurial entry on competing crowdfunding platforms. *Strategic Management Journal*, *19*(3). https://sms.onlinelibrary.wiley.com/journal/1932443x

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, *10*(28), Article eadn5290. https://doi.org/10.1126/sciadv.adn5290

George, M. M., Strauss, K., Mell, J. N., & Vough, H. C. (2023). When "who I am" is under threat: Measures of threat to identity value, meanings, and enactment. *Journal of Applied Psychology*, *108*(12), 1952–1978. https://doi.org/10.1037/apl0001114

Gonzalez, X. (2023, August 24). M.F.A. vs. GPT. *The Atlantic*. https://www.theatlantic.com/ideas/archive/2023/08/mfa-chat-gpt-future/675090/

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), Article 619. https://doi.org/10.1126/science.1134475

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press. https://doi.org/10.1111/jedm.12050

Hitsuwari, J., Ueda, Y., Yun, W., & Nomura, M. (2023). Does human–AI collaboration lead to more creative art? Aesthetic evaluation of human-made and AI-generated haiku poetry. *Computers in Human Behavior*, *139*, Article 107502. https://doi.org/10.1016/j.chb.2022.107502

Ho, C.-C., & MacDorman, K. (2017). Measuring the uncanny valley effect: Refinements to indices for perceived humanness, attractiveness, and eeriness. *International Journal of Social Robotics*, *9*(1), 129–139. https://doi.org/10.1007/s12369-016-0380-9

Hong, J.-W., Fischer, K., Ha, Y., & Zeng, Y. (2022). Human, I wrote a song for you: An experiment testing the influence of machines' attributes on the AI-composed music evaluation. *Computers in Human Behavior*, *131*, Article 107239. https://doi.org/10.1016/j.chb.2022.107239

Horton, C. B., Jr., White, M. W., & Iyengar, S. S. (2023). Bias against AI art can enhance perceptions of human creativity. *Scientific Reports*, *13*(1), Article 19001. https://doi.org/10.1038/s41598-023-45202-3

Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, *14*(1), Article 3440. https://doi.org/10.1038/s41598-024-53303-w

Jago, A. S. (2019). Algorithms and authenticity. *Academy of Management Discoveries*, *5*(1), 38–56. https://doi.org/10.5465/amd.2017.0002

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior*, *114*, Article 106553. https://doi.org/10.1016/j.chb.2020.106553

Liu, B., & Sundar, S. S. (2018). Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking*, *21*(10), 625–636. https://doi.org/10.1089/cyber.2018.0110

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing*, *86*(1), 91–108. https://doi.org/10.1177/0022242920957347

Lucas, B. J., & Nordgren, L. F. (2020). The creative cliff illusion. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(33), 19830–19836. https://doi.org/10.1073/pnas.2005620117

Martin, A. E., & Mason, M. F. (2022). What does it mean to be (seen as) human? The importance of gender in humanization. *Journal of Personality and Social Psychology*, *123*(2), 292–315. https://doi.org/10.1037/pspa0000293

Mellamphy, N. B. (2021). Humans "in the loop"? *Nature and Culture*, *16*(1), 11–27. https://doi.org/10.3167/nc.2020.160102

Mineo, L. (2023, August 15). Is art generated by artificial intelligence real art? *The Harvard Gazette*. https://news.harvard.edu/gazette/story/2023/08/is-art-generated-by-artificial-intelligence-real-art/

Newson, R. (2022). *PARMEST: Stata module to create new data set with one observation per parameter of most recent model* [Computer software]. Statistical Software Components. https://ideas.repec.org//c/boc/bocode/s352601.html

Ornes, S. (2019). Science and culture: Computers take art in new directions, challenging the meaning of "creativity". *Proceedings of the National Academy of Sciences of the United States of America*, *116*(11), 4760–4763. https://doi.org/10.1073/pnas.1900883116

Park, J., Javalgi, R., & Wachter, M. (2016). Product ethnicity and perceived consumer authenticity: The moderating role of product type. *Journal of Consumer Marketing*, *33*(6), 458–468. https://doi.org/10.1108/JCM-01-2015-1272

Rogers, R. (2024, December 11). Generative AI is my research and writing partner. Should I disclose it? *Wired*. https://www.wired.com/story/prompt-disclose-at-in-creative-work-teach-kids-about-chatbots/

Roose, K. (2022, September 2). An A.I.-generated picture won an art prize. Artists aren't happy. *The New York Times*. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

Schanke, S., Burtch, G., & Ray, G. (2021). Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research*, *32*(3), 736–751. https://doi.org/10.1287/isre.2021.1015

Shaffi, S. (2023, January 23). 'It's the opposite of art': Why illustrators are furious about AI. *The Guardian*. https://www.theguardian.com/artanddesign/2023/jan/23/its-the-opposite-of-art-why-illustrators-are-furious-about-ai

Shah, A. (2021). *ASDOC: Stata module to create high-quality tables in MS word from Stata output* [Computer software]. Statistical Software Components. https://ideas.repec.org//c/boc/bocode/s458466.html

Shank, D. B. (2025). *The machine penalty: The consequences of seeing artificial intelligence as less than human*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-86061-4

Shank, D. B., Stefanik, C., Stuhlsatz, C., Kacirek, K., & Belfi, A. M. (2023). AI composer bias: Listeners like music less when they think it was composed by an AI. *Journal of Experimental Psychology: Applied*, *29*(3), 676–692. https://doi.org/10.1037/xap0000447

Sherrer, K. (2025, March 17). Should AI write fiction? OpenAI did & an author called it "pastiche garbage". *eWEEK*. https://www.eweek.com/news/openai-sam-altman-ai-creative-writing-authors/

StataCorp. (2021). *STATA/MP* (Version 17) [Computer software]. https://www.stata.com/

Thompson, M. M., Zanna, M. P., & Griffin, D. W. (1995). Let's not be indifferent about (attitudinal) ambivalence. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 361–386). Lawrence Erlbaum.

Ticong, L. (2025, February 28). Oscars set new AI disclosure as generative tech creeps into filmmaking. *eWEEK*. https://www.eweek.com/news/oscars-ai-disclosure-filmmaking/

Tigre Moura, F., Castrucci, C., & Hindley, C. (2023). Artificial intelligence creates art? An experimental investigation of value and creativity perceptions. *The Journal of Creative Behavior*, *57*(4), 534–549. https://doi.org/10.1002/jocb.600

Tong, S., Jia, N., Luo, X., & Fang, Z. (2021). The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, *42*(9), 1600–1631. https://doi.org/10.1002/smj.3322

Walt, V. (2023, December 27). What's next in artificial intelligence? *The New York Times*. https://www.nytimes.com/2023/12/27/business/dealbook/artificial-intelligence-investment-laws.html

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(43), 11374–11379. https://doi.org/10.1073/pnas.1704347114

Wu, Y., Mou, Y., Li, Z., & Xu, K. (2020). Investigating American and Chinese subjects' explicit and implicit perceptions of AI-generated artistic work. *Computers in Human Behavior*, *104*, Article 106186. https://doi.org/10.1016/j.chb.2019.106186

Zhou, E., & Lee, D. (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus*, *3*(3), Article pgae052. https://doi.org/10.1093/pnasnexus/pgae052

Received February 24, 2025
Revision received October 8, 2025
Accepted October 27, 2025 ■