The Agency Trade-Off: Generative AI Reshapes Two Dimensions of Human Agency in Opposite Directions

Benjamin Lira Luttges[1,*], Angela Duckworth[2], Lyle Ungar[3]

[1]The Wharton School, University of Pennsylvania. [2]Department of Psychology, University of Pennsylvania. [3]Department of Computer and Information Science, University of Pennsylvania.

[*]Corresponding author: blira@upenn.edu.

Word count: ∼6,700 (main text including Methods).

Abstract. Does using generative artificial intelligence (AI) enhance or erode human agency? We argue it does both, on different dimensions. A factor-analytic study (N = 399 U.S. adults reporting on a recent AI-assisted writing task) shows that two facets of agency are psychometrically distinct and carry different downstream correlates: **outcome agency**, felt because we *achieve* a goal, and **process agency**, felt because we *control how* we act toward it. We then run two preregistered studies on nationally representative data collected in partnership with the Gallup organization (N = 3627, N = 1533). In a 2 × 2 vignette experiment, respondents predicted that a writer using AI would lose more than a full standard deviation of process agency (d = 1.38), and that even his successes would feel less meaningful because they credited his wins to the AI rather than to him. A separate cohort in the same panel wrote captions for a *New Yorker*-style cartoon with or without AI. Their process agency dropped sharply (d = -1.18) while their outcome agency instead *rose* (d = +0.47), and their meaning ratings matched those of writers who worked alone. AI's net effect on meaning therefore depends on which side of this trade a given task, tool, and construal level emphasize.

People increasingly delegate work to AI. Side-by-side experiments show that handing off writing, code, analysis, and design to a generative AI assistant produces faster output, often at quality the user could not have reached alone[1–3]. This has been accompanied by concern that working with AI reduces agency. In a 2025 nationally representative survey of U.S. young adults, 79% endorsed the view that AI is making people lazier and less capable[4], even though most of those same respondents use AI tools themselves (74% in the prior 30 days). Survey and ethnographic work on AI users splits along similar lines: some say AI extends what they can do, others say it has taken something they used to do well[5,6].

The implicit model behind the worry treats agency as a single quantity that AI depletes: the more AI does something, the less involved you are in it. On that model, productivity gains come at a one-to-one cost to agency. The loss would matter, because agency anchors the meaning people draw from their work[7] and predicts motivation, well-being, and persistence on hard tasks[8–12].

But agency is not a single quantity. We track two facets separately for AI assistance. The first is **outcome agency**: the agency a person feels because they *brought about the result they wanted*. The second is **process agency**: the agency a person feels because they controlled the *process of how things were done*. The distinction runs the length of a goal hierarchy, in which superordinate ends sit atop the subordinate means that achieve them[13–17]: outcome agency is felt control at the level of the end, process agency at the level of the means. The two are conceptually, empirically, and phenomenologically distinct[8,18]. The prediction this paper tests is that AI moves the two in opposite directions.

Normally, the two coincide. To bring about a result you want, you have to do the work, so the means and the end stay fused: doing the writing and owning its effect on the reader are one experience[19,20]. The fusion is reliable enough in ordinary life that the two facets rarely come apart. AI breaks it, delivering the end without requiring the means.

AI raises outcome agency by producing better results than the user would have reached alone—the writer who pushes the AI button finishes a stronger draft. And it erodes process agency by leaving on the page words and structure the user can no longer recognize as their own.

It follows that "does AI cost us agency?" is the wrong question. The right question is which facet of agency AI moves. Because both facets feed into the meaning people draw from their work, AI affects meaning through two channels at once: one that lifts meaning (the work turned out better than it would have alone) and one that drags meaning down (the work was less recognizably yours). The two can cancel.

We test the account in three studies. Study 1 (N = 399) establishes the distinction: process agency and outcome agency factor cleanly into two dimensions in a U.S. adult sample of recent AI users, and both predict meaning. Study 2 (N = 3627), a 2 × 2 vignette experiment embedded in a nationally representative panel we collected in partnership with

the Gallup organization, asks respondents to rate the agency of a writer named Alex who composed a petition either with ChatGPT or without and whose petition either succeeded or failed. Respondents rated an Alex who used AI as having more than a full standard deviation less process agency than a solo Alex, judged even his successes as less meaningful, and transferred credit for his wins to the AI. Study 3 (N = 1533) gives a separate cohort in the same panel the actual experience: they wrote captions for a *New Yorker*-style cartoon with or without AI assistance. Using AI simultaneously lowered their process agency and raised their outcome agency. Using AI did not enhance or reduce meaning, because its positive effect on performance and its negative effect on originality offset.

Process and outcome agency are psychometrically distinct

We ran a factor-analytic study on naturally occurring instances of AI-assisted writing. 399 U.S. adults recruited through Prolific—67% of whom had used ChatGPT, 15% Gemini, and 9% Microsoft Copilot for the experience they described—had used an AI tool for writing in the prior month and were asked to recall a specific recent experience in which they had used AI to help with a writing task. They described the task in their own words (e.g., "cover letter writing," "a list of brainstorm ideas," "writing marketing copy for a new product," "writing a summary of a book or article") and reported what fraction of the work the AI did. Most respondents were employed full-time (65%) or part-time/self-employed (27%), and the recalled tasks ranged across personal writing (e.g., a message to someone on a dating app), hobby projects (e.g., a newsletter for a cat sanctuary), and work-related writing (cover letters, emails to customers, marketing copy, code summaries, process documents). On average, participants estimated that AI did 5 of every 10 units of work; the modal response indicated a roughly even split between user and tool.

Participants then rated four process-agency items (e.g., "I felt like I was doing the work"), four outcome-agency items (e.g., "I felt effective"), and three meaning items (proud, meaningful, purposeful) on 0–10 scales. A separate three-item composite measured the writer's *writing identity*—how important writing is to who they are (α = 0.93; M = 5.84, SD = 2.66). We fit a two-factor confirmatory factor analysis using robust maximum likelihood and compared its fit against a one-factor alternative via the Satorra-Bentler scaled $\Delta\chi^2$ test.

A two-factor model fit the data well and outperformed a one-factor alternative. The two-factor solution fit well (CFI = 0.944, TLI = 0.918, RMSEA = 0.138, SRMR = 0.042) and significantly outperformed a one-factor solution (CFI = 0.860, $\Delta\chi^2$ = 40.8, $p < .001$). The two factors correlated r = 0.83, indicating that they are related but distinct. Both composites were internally consistent (α_{PA} = 0.90; α_{OA} = 0.91; Table 1).

Table 1. Study 1 standardized loadings from the two-factor confirmatory model.

Item	λ
Process agency	
I felt in control of the process	0.81
I felt like I was doing the work	0.82
I felt responsible for how the work was done	0.83
I felt involved in the process	0.89
Outcome agency	
I felt in control of the outcome	0.88
I felt ownership over the results	0.88
I felt responsible for the outcome	0.87
I felt effective	0.79

The two scales had different downstream correlates. Using Steiger's Z for dependent correlations, outcome agency correlated more strongly with felt outcome credit (r = 0.72) than process agency did (r = 0.55; Δr = +.17, Z = -6.58, $p < .001$), and outcome agency correlated more strongly with outcome satisfaction (r = 0.65) than process agency did (r = 0.54; Δr = +.11, Z = -3.85, $p < .001$). Process agency correlated numerically more strongly with meaning (r = 0.60) than outcome agency did (r = 0.54), although this difference fell short of conventional significance (Z = 1.85, $p = .064$). Negative attitudes about AI's effect on users—a 4-item composite of agreement with "AI chatbots make people less smart / lazier / less learning over time / less happy"—correlated equally with both agency dimensions (r = -0.37 vs. -0.37; Z = -0.01, $p = .993$), suggesting that broad concerns about AI's effect on users act on agency as a whole rather than on either facet (Supplement SS4).

Writing identity moderated the process-agency–meaning link. We tested whether the strength of each agency dimension's link to meaning depends on how central writing is to the person. Writing identity moderated the process-agency–meaning relationship ($b_{PA\times ID}$ = 0.037, $p = .011$). The slope linking process agency to meaning was substantial for writers low in writing identity (slope = 0.32, $p < .001$) and roughly 60% steeper for writers high in writing identity (slope = 0.52, $p < .001$; Figure 1). Writing identity did not moderate the outcome-agency–meaning relationship ($b_{OA\times ID}$ = 0.013, $p = .429$); the slope was comparable at low writing identity (0.23, $p = .006$) and high writing identity (0.30, $p < .001$). For writers whose identity is invested in the writing, process agency carries more weight in how meaningful the experience feels; for outcome agency, identity makes little difference.

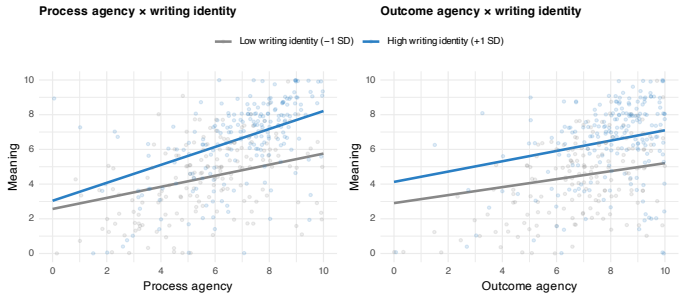

Fig. 1. Study 1: writing identity moderates the link between process agency and meaning (left panel) but not the link between outcome agency and meaning (right panel). Predicted lines from regression models with mean-centered predictors; raw data jittered for visibility.

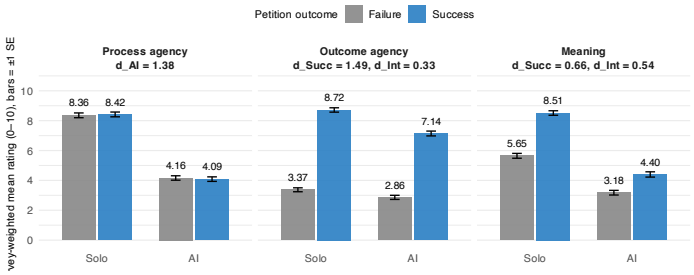

Fig. 2. Study 2: Alex ratings by tool and petition outcome (N = 3627). Bars show condition means; error bars are ±1 standard error. In-panel labels show preregistered effect sizes: process agency reports the AI main effect; outcome agency and meaning report the success main effect and the AI × success interaction.

When asked to imagine an AI user, people predict an across-the-board agency loss

Study 1 showed that process agency and outcome agency are separable. Study 2 asks how people expect AI assistance to act on each. We embedded a preregistered 2 × 2 vignette experiment (AsPredicted #288,196) in a survey we ran in partnership with the Gallup organization (3627 respondents; ages 18–83; Gen Z adults and the adult parent respondents in the household sample). Respondents read a short scenario about a writer named Alex composing a petition to increase library funding. We orthogonally varied whether Alex used ChatGPT or Microsoft Word and whether the petition succeeded or failed. Respondents then rated Alex on three process-agency, three outcome-agency, and three meaning items on 0–10 scales (Cronbach's α = 0.96, 0.92, 0.95). Cell means and standard deviations for all four condition cells appear in Supplement S3.

Respondents who imagined Alex using AI rated his authorship of the writing far lower than respondents who imagined him writing alone. On the 0–10 process-agency scale, an Alex who used Microsoft Word averaged 8.36–8.42, while an Alex who used ChatGPT averaged 4.16–4.09. The contrast corresponds to d = 1.38 ($p < .001$; Figure 2). The size of the cut was roughly the same whether the petition succeeded or failed; success did not move how respondents rated who had authored the writing.

For outcome agency, the dominant signal was whether Alex succeeded; AI assistance produced a smaller but consistent secondary penalty. Outcome agency was driven primarily by whether Alex's petition succeeded (d = 1.49, $p < .001$): a successful Alex was rated much more capable than a failed Alex regardless of tool. On top of that, an Alex who used AI was rated slightly less capable than a solo Alex (d = 0.30, $p = .008$). The AI penalty grew larger when the petition succeeded than when it failed: the AI × success interaction was significant (b = -1.07, $p < .001$; d_{Int} = 0.33), indicating that respondents transferred some of the credit for a successful petition from Alex to the AI rather than docking him uniformly for using the tool.

Meaning showed the same structure: success drove the main effect, with AI suppressing the credit for the upside. Success raised meaning ratings by d = 0.66 ($p < .001$), but a successful Alex who used AI was judged considerably less meaningful than a successful solo Alex (AI × success interaction b = -1.64, $p < .001$; d_{Int} = 0.54). Across process, outcome, and meaning, respondents directed the largest agency penalty at the part of the experience AI had most visibly touched (the writing itself) and a smaller, structurally similar penalty at the credit they were willing to give Alex for the result.

Younger respondents were the harshest judges. We preregistered a test of whether the AI penalty varied by respondent age. It did. Translating the preregistered interaction terms ($b_{\text{AI}\times\text{age}}$ = 0.034, $p = .015$ for process agency; 0.034, $p = .033$ for meaning) into simple slopes at ±1 SD of age (SD = 10.5 years): respondents one SD below the mean age cut Alex's process agency by b = -4.63 ($p < .001$) on the 0–10 scale, while respondents one SD above the mean age cut it by b = -3.91 ($p < .001$)—a roughly one-point softening across the age range. The pattern was similar for meaning (younger: b = -3.64; older: b = -2.94).

When people actually use AI themselves, they experience a trade rather than an across-the-board loss

Study 2 captures what respondents predict an AI user would experience. Study 3 asks what an AI user actually reports having experienced. In a preregistered behavioral experiment (AsPredicted #288,197) in the same Gallup panel (N = 1533; ages 12–29; 787 solo, 746 AI), participants wrote a funny caption for a *New Yorker*-style cartoon depicting penguins overrunning a city sidewalk (Figure 3). In the AI condition, a single button press generated five caption suggestions for

Fig. 3. The cartoon stimulus used in Study 3 (and in the pilot cartoon waves reported in the supplement). All participants saw the same image and were asked to write a funny caption. Cartoon by Meredith Southard.

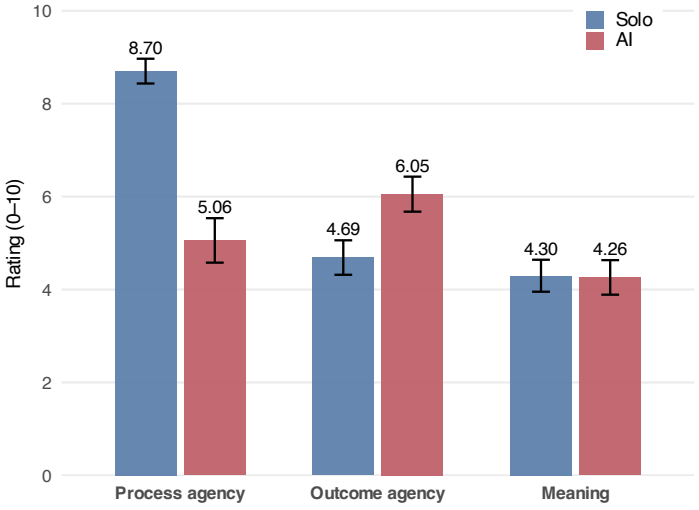

Fig. 4. Study 3: process agency, outcome agency, and meaning by condition in the Gallup cartoon experiment (N = 1533). Bars show means ± 95% CI from independent participant subsamples (each participant rated one rating block).

the same image, each accompanied by an objective humor score from a model we fine-tuned for the study. Participants could submit any suggestion verbatim, edit it, or write their own; they were required to press the AI button at least once. After submission, all participants received a single feedback indicator (a thumbs-up or thumbs-down) generated from the pretrained humor model's evaluation of their caption.

Each participant was then randomly assigned to rate only one of the three dependent variables: process agency (513 participants), outcome agency (474), or meaning (546). We adopted this single-rating-block design after documenting an order-of-measurement artifact in three earlier Prolific pilots in which every participant rated all three dependent variables in a counterbalanced order. In those pilots, rating process agency *first* systematically lowered subsequent ratings of outcome agency and meaning in the AI condition—in some waves it flipped the direction of the AI effect on those downstream measures. Drawing the user's attention to the part of their experience AI had most touched (the writing itself) appeared to act as a manipulation in its own right. Randomizing each participant to a single rating block sidesteps this contamination; the full per-position analysis is in Supplement SS6. All composites were highly reliable (α = 0.96, 0.94, 0.94).

Participants who wrote with AI reported much lower authorship of the work than participants who wrote alone. On the 0–10 process-agency scale, participants who wrote alone averaged M = 8.70, while participants who wrote with the AI tool averaged M = 5.06 (d = -1.18, $p < .001$). The direction matches the prediction respondents in Study 2 made; the magnitude is roughly half what they predicted.

The same participants rated themselves as more capable of producing a funny caption than participants who wrote alone. On the 0–10 outcome-agency scale, participants who wrote with AI averaged M = 6.05 versus M = 4.69 for participants who wrote alone (d = +0.47, $p < .001$). This is directionally opposite the prediction respondents in Study 2 made.

Meaning did not differ overall between the two groups. On the 0–10 meaning scale, participants who wrote with AI averaged M = 4.26 compared with M = 4.30 for participants who wrote alone (d = -0.01, $p = .888$).

The outcome-agency difference covaried with objective caption quality and the feedback participants received. Captions written with AI scored substantially higher on the pretrained humor model than captions written alone (Solo M = 37.0, AI M = 60.8; d = 0.98, $p < .001$). Participants who wrote with AI were also more likely to receive a thumbs-up than participants who wrote alone (69% vs. 29%; OR = 5.38, 95% CI [4.31, 6.75]; $\chi^2(1) = 241.8$, $p < .001$). A standardized parallel-mediation model (Figure 5; full estimates in Supplement SS7) finds that the standardized association between AI condition and outcome agency, controlling for log keystrokes and caption–AI similarity, runs predominantly through the thumbs-up indicator (indirect via feedback: b = 0.32, $p < .001$). We report these paths as descriptive of the joint distribution rather than as causal estimates; the mediators were not independently manipulated.

The trade-off: similarity to AI predicts process agency and caption quality in opposite directions. Within the AI condition, the maximum word-Jaccard similarity between each

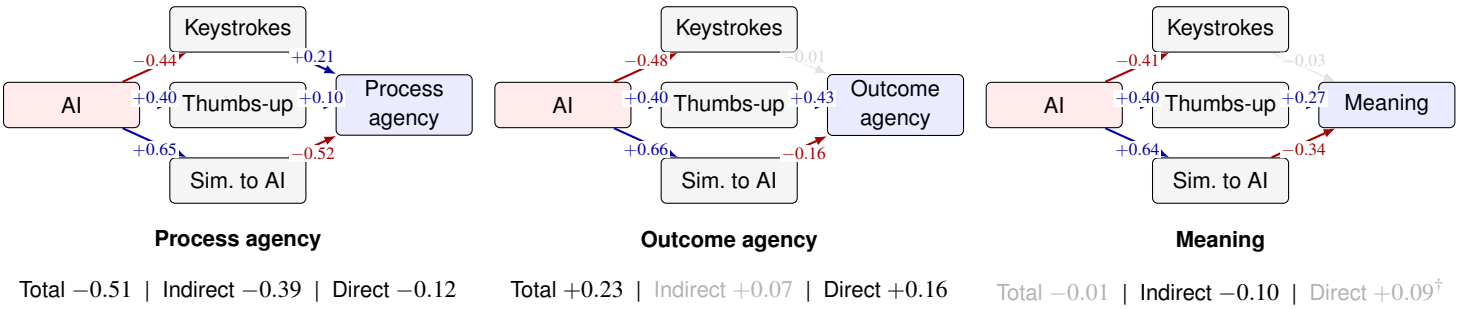

Fig. 5. Study 3 parallel-mediation path estimates linking AI condition to each agency dimension through three mediators. All coefficients are standardized; solid colored arrows indicate $p < .05$ (blue = positive, red = negative); faded gray arrows are non-significant. Numbers beneath each panel report the standardized total effect of AI, the total indirect effect through all three mediators, and the direct effect of AI controlling for the three mediators († marks $p < .10$). We report these as descriptive of the joint distribution rather than as causal estimates: the mediators were not independently manipulated.

participant's submitted caption and any of the AI suggestions they had seen was strongly diagnostic of where on the trade-off they landed. The more of the AI's wording a participant adopted, the less process agency they reported ($r = -0.69$, $p < .001$; Figure 6A) and the higher the pretrained humor score on their submission ($r = +0.49$, $p < .001$; Figure 6B). Sample captions illustrate the two ends. At the low-similarity end—participants who used the AI as a springboard and wrote their own caption—submissions included *"Holy penguin"* and *"This neighborhood has really changed since we first moved here."* At the high-similarity end—participants who submitted an AI suggestion essentially verbatim—submissions included *"Rush hour in Antarctica sure looks different this year!"* (Jaccard = 1.0 with an AI suggestion).

The right-panel intercept makes the joint story explicit. Even at the lowest-similarity end of the AI condition, the smoothed humor score sits above the Solo mean, so adopting little or no AI text still produced funnier captions than writing alone. But the smoother crosses the Solo authorship line at high similarity in the left panel: participants who copied AI text wholesale gave up most of their authorship. The more independently a participant wrote with AI in hand, the more of their process agency they protected and the smaller the head start on humor they ended up with.

For meaning, the two indirect paths offset. The indirect path through feedback was positive ($b = 0.21$, $p < .001$); the indirect path through caption–AI similarity was negative ($b = -0.43$, $p < .001$); the total indirect effect was modestly negative ($b = -0.20$, $p = .008$), and the direct effect was positive, leaving the simple condition difference indistinguishable from zero.

Younger and older participants reported the same trade. The age-graded harshness we saw when respondents in Study 2 imagined an AI user did not appear when respondents in the same panel actually wrote with AI themselves. Adolescents reported about the same drop in authorship and about the same gain in capability as young adults did. Details and exploratory moderation models are reported in Supplement SS7.

Discussion

Three findings converge. Process agency and outcome agency are psychometrically distinct. When adolescents and young adults in the Gallup panel actually used AI to write a caption, their process agency dropped by more than a full standard deviation while their outcome agency rose by half of one. And when respondents in the same panel were instead asked to imagine an AI user, they predicted that he would lose agency on both dimensions at once.

The gap between what people predict and what users report has practical consequences for adoption. A young respondent who believes, as Study 2 respondents did, that using AI on a creative writing task will gut both his sense of capability *and* his sense of ownership has good reason to avoid the tool. Study 3 participants in the same panel did pay a real cost: they reported a process-agency loss of more than a full standard deviation, a loss that identity-invested writers (per Study 1) feel especially keenly. What they did *not* pay was the second cost respondents expected—a matching drop in outcome agency—and the small mean gain in outcome agency they did report came primarily from receiving positive feedback on a higher-scoring caption. Whether the net of these moves counts as a worthwhile trade depends on how much a particular person values the authorship they gave up relative to the success they received. But Study 3 participants—in the same panel, facing the same kind of identity-expressive task—came away feeling more capable, not less. Underadoption may follow from an inaccurate prediction of what using the tool actually feels like.

A second design implication follows from where in the chain from intention to outcome AI does its work. Classical accounts of agency treat the experience as built on the coher-

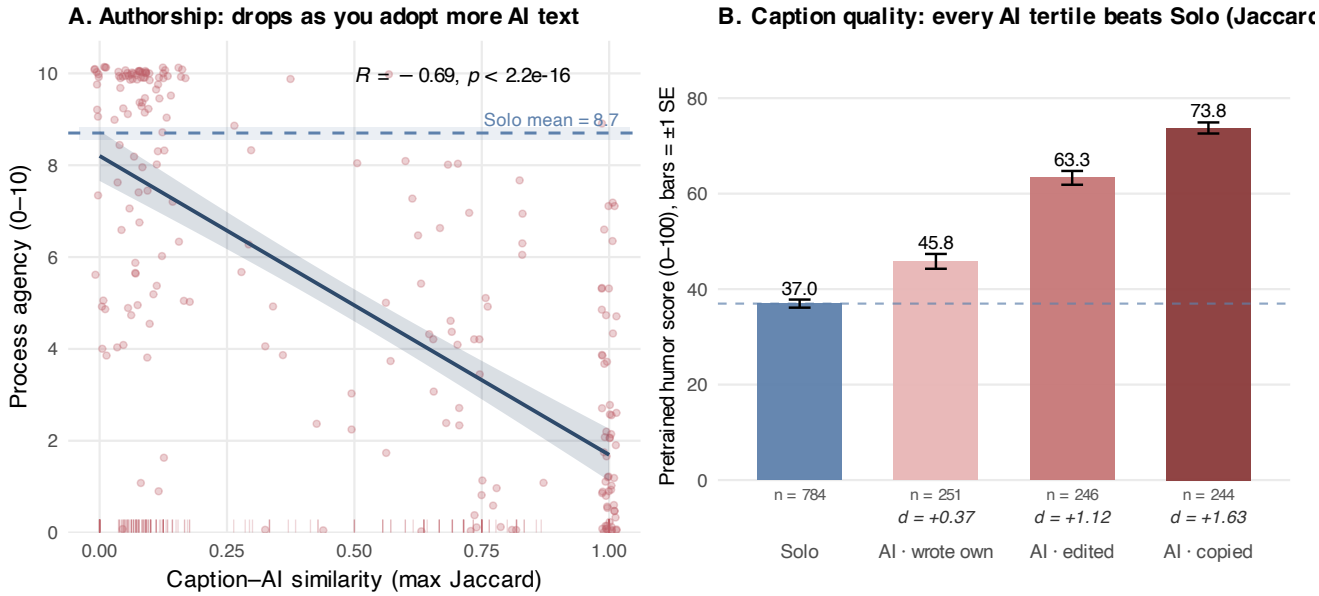

Fig. 6. Study 3 trade-off, AI condition only. Each point is one AI-condition participant; the x-axis is the maximum word-Jaccard similarity between that participant's submitted caption and any of the AI suggestions they had seen. (A) Process agency drops as participants adopt more AI text. (B) Pretrained humor score rises as participants adopt more AI text. The dashed horizontal line in each panel marks the Solo mean (with a ± 1 SE band); rug ticks at the bottom show the marginal distribution of similarity. Across all three Jaccard tertiles, the AI condition outperforms Solo on humor (Supplement SS8 reports tertile bars and a continuous cutoff sweep, plus an embedding-cosine robustness check).

ence of three steps—a person forms an intention, executes an action, and observes the outcome of that action[8,18,21]. AI assistance keeps the first and third steps with the user (the user still sets the goal and submits the work) and substitutes for the second. The longer and more visible the user's contribution to the action step, the more process agency survives. Tools that require multiple decisions, edits, or selections from the user should leave more authorship in place than tools that compress the action into a single "generate" button.

Our findings give designers a concrete handle on how to build AI tools that preserve process agency. Outcome agency tracks whether the work succeeded, so any AI tool that improves performance will raise it on average. Process agency, in contrast, tracks how much of the visible product the user can recognize as their own—the more the user wrote, edited, or shaped, the more authorship remains theirs. Tools that surface options, require choices, or scaffold editing should leave more process agency in place than tools that auto-complete entire passages. The process loss is not a fixed cost of AI use; it varies with the specific shape of the human contribution the tool leaves room for.

We are skeptical, however, that design choice will close the gap on its own. The incentives that surround knowledge work—deadlines, quotas, billable hours, evaluations that reward output and not the process behind it—reward delegation rather than scaffolded use. A student facing a paper deadline, an analyst facing a midnight memo, and a marketer facing a Monday-morning launch all face the same local pressure: take the version the AI generates and ship it. The agency-preserving design choices we describe are real, but using them takes time the user is being asked to save. Whether process-preserving tools are adopted in the workplaces, schools, and side projects where AI use is growing fastest depends on whether the people using them are evaluated on the work they did, or only on the work they delivered.

AI's null effect on meaning in Study 3 hides two opposing forces. Among AI users, receiving positive feedback raised meaning ratings; submitting an AI-like caption lowered them. Meaning is not an independent third construct but the net of the two agency dimensions, with the relative weights varying across people and tasks. This view aligns with the IKEA-effect literature[22]: attachment and felt meaning attach to what people have themselves made, so the meaning a piece of work carries depends in part on how much of it the worker can recognize as her own. Study 1's writing-identity moderation is consistent: writers whose identity was bound up in the activity showed a steeper association between process agency and meaning, while the outcome-agency–meaning association did not vary with identity. Practically, this suggests the same process-agency loss may register more strongly on meaning for users whose identity is invested in the work (a novelist handing over her prose) than for users for whom the work is instrumental (an accountant handing over a routine

email), even at the same drop in authorship.

Limitations. Two limitations point to directions for follow-up work. First, Study 3 captures a single brief creative task. The trade-off may take different forms in domains where the process *is* the product rather than the means to it—the proof in mathematics, the performance in music, the move-by-move play in chess—where the audience values the unfolding of the doing, not just the result, and substituting AI for the doing may not preserve what makes the artifact valuable. The trade-off may also take different forms in domains where the outcome is not legible to the user themselves: an open-ended research question for which there is no graded answer key, a long parenting decision whose effects unfold over years, a draft of a novel that no one reads for a year. When the user cannot tell whether the outcome was good, only the process is available as a signal of self-cause, and AI's substitution for that process should weigh more. Tasks that consumers perceive as more subjective[23] are a natural candidate for sharper process-agency losses than the modal task in our cartoon study and warrant a direct test. Second, all three studies measure a single brief exposure; sustained AI use over weeks or months may produce different patterns, including cumulative effects that single-shot designs cannot capture. A related concern that our individual-level designs are not built to detect is the collective homogenization of human creative output under widespread AI use[24,25]; the gains on individual outcome agency we document here may coexist with losses to the aggregate diversity of human work.

Conclusion. The worry that AI is eroding human agency is real, and our data suggest it is right on one dimension: AI substantively cuts into the experience of doing the work. But the worry is wrong on the other: the same AI users feel *more* capable of producing the result, not less. The mistake worried observers have been making—and that the field made too, when it treated AI's effect on agency as a single question—is to collapse outcome agency and process agency into one. Once the two are separated, the question of what AI does to human agency stops being a single up-or-down referendum and becomes a pair of concrete design questions: which AI tools preserve the user's hand in the work, and which incentive structures reward the doing rather than only the delivery? Whether AI ends up enhancing or eroding human agency at scale depends on how those questions get answered.

Methods

Study 1: Factor analysis

Participants were 399 U.S. adults recruited through Prolific who reported using AI tools for writing in the prior 30 days. They were asked to recall a specific recent experience and rate four process-agency items and four outcome-agency items (Table 1) on 0–10 scales, along with single-item measures of meaning and outcome satisfaction and a three-item writing-identity composite. We fit a two-factor confirmatory factor analysis using robust maximum likelihood (`lavaan`; MLR estimator) and compared its fit to a one-factor alternative via the Satorra-Bentler scaled $\Delta\chi^2$ test.

Study 2: Gallup vignette

We embedded a 2 (tool: ChatGPT vs. Microsoft Word) $\times$ 2 (outcome: success vs. failure) between-subjects vignette in a survey administered in partnership with the Gallup organization (3627 respondents; ages 18–83). The design was preregistered (AsPredicted #288,196). Respondents read a short scenario about Alex composing a library-funding petition and observing the petition's outcome, then rated Alex on three process-agency, three outcome-agency, and three meaning items on 0–10 scales (Cronbach's α = 0.96, 0.92, 0.95). For each dependent variable, we fit a linear model with main effects of AI condition and petition outcome and their interaction. A preregistered secondary analysis added respondent age, mean-centered, and its interaction with AI condition; the corresponding coefficient indexes whether the AI effect varies linearly with respondent age.

Study 3: Gallup cartoon experiment

We ran a 2 (condition: solo vs. AI) $\times$ 3 (rating block: process agency / outcome agency / meaning) design as an add-on to the same Gallup panel; the design was preregistered (AsPredicted #288,197). Each participant was randomly assigned to rate one of the three dependent variables. Participants in the AI condition saw a button labeled *"See some suggestions"*; each press produced five GPT-4o caption suggestions for the same image, each accompanied by a 0–100 humor score from a pretrained scoring model. After submission, all participants received a single feedback indicator (a thumbs-up if their caption's pretrained humor score landed above the solo median, a thumbs-down otherwise) before completing their assigned rating block. Each block contained three items on a 0–10 scale (process: producing, responsible, involved; outcome: capably, effectively, successfully produced a funny caption; meaning: fulfilling, meaningful, mattered). The preregistered primary test was a linear model regressing the assigned dependent variable on condition; we report Cohen's d alongside.

Humor scoring model

To give participants timely feedback that did not rely on crowd-sourced voting within the study itself, we built an objective caption-scoring model. Specifically, we fine-tuned GPT-4.1-mini on human pairwise judgments collected in earlier waves of the project: we fit a Bradley-Terry model to 14,350 pairwise votes across 729 captions, converted log-strengths to percentile ranks against the solo distribution,

and trained the language model to predict these calibrated 0–100 scores from caption text. On a held-out validation set (n = 146), the model achieved Pearson r = .62 between predicted and actual scores, with mean absolute error of 17.8 points and 75% accuracy classifying captions into the top vs. bottom half of the solo distribution. The same model also scored each AI suggestion displayed to participants in the AI condition.

Caption–AI similarity

For each AI-condition participant, we computed the maximum word-level Jaccard similarity between their submitted caption and any of the AI suggestions they had seen. Solo participants saw no AI text and were assigned similarity = 0.

Parallel mediation

For each dependent variable in Study 3 we fit a parallel-mediation structural model in `lavaan` with three standardized mediators (log keystrokes, thumbs-up indicator, max AI similarity), a direct AI $\rightarrow y$ path, and 2,000 bootstrap resamples for percentile confidence intervals on the indirect effects.

Open science

Preregistrations, materials, analysis code, and de-identified data are posted at the project repository. Studies were approved by the University of Pennsylvania IRB.

1. S. Noy, W. Zhang, Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187–192 (2023).
2. E. Brynjolfsson, D. Li, L. R. Raymond, Generative AI at work. *NBER Working Paper* (2023).
3. F. Dell'Acqua, E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, K. R. Lakhani, Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality, *Working Paper 24-013*, Harvard Business School (2023).
4. B. Lira, L. Ungar, A. L. Duckworth, Coach, not crutch: What generative AI is really doing to learning (2026).
5. N. Vincent, H. Li, N. Tilly, B. Hecht, *Working paper* (2024).
6. A. Khurana, H. Subramonyam, P. K. Chilana, *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems* (2024).
7. J. R. Hackman, G. R. Oldham, Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance* **16**, 250–279 (1976).
8. P. Haggard, Sense of agency in the human brain. *Nature Reviews Neuroscience* **18**, 196–207 (2017).
9. A. Bandura, *Self-efficacy: The exercise of control* (W.H. Freeman, New York, 1997).
10. A. Tapal, E. Oren, R. Dar, B. Eitam, The Sense of Agency Scale: A Measure of Consciously Perceived Control over One's Mind, Body, and the Immediate Environment. *Frontiers in Psychology* **8**, 1552 (2017).
11. E. L. Deci, R. M. Ryan, The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry* **11**, 227–268 (2000).
12. R. M. Ryan, E. L. Deci, Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* **55**, 68–78 (2000).
13. C. S. Carver, M. F. Scheier, Control theory: A useful conceptual framework for personality-social, clinical, and health psychology. *Psychological Bulletin* **92**, 111–135 (1982).
14. C. S. Carver, M. F. Scheier, *On the self-regulation of behavior* (Cambridge University Press, New York, 1998).
15. A. W. Kruglanski, J. Y. Shah, A. Fishbach, R. Friedman, W. Y. Chun, D. Sleeth-Keppler, A theory of goal systems. *Advances in Experimental Social Psychology* **34**, 331–378 (2002).
16. R. R. Vallacher, D. M. Wegner, What do people think they're doing? Action identification and human behavior. *Psychological Review* **94**, 3–15 (1987).
17. Y. Trope, N. Liberman, Construal-level theory of psychological distance. *Psychological Review* **117**, 440–463 (2010).
18. E. Pacherie, The phenomenology of action: A conceptual framework. *Cognition* **107**, 179–217 (2008).
19. A. W. Kruglanski, A. Fishbach, K. Woolley, J. J. Bélanger, M. Chernikova, E. Molinario, A. Pierro, A structural model of intrinsic motivation: On the psychology of means-ends fusion. *Psychological Review* **125**, 165–182 (2018).
20. A. Fishbach, M. J. Ferguson, The goal construct in social psychology. *Social Psychology: Handbook of Basic Principles*, A. W. Kruglanski, E. T. Higgins, eds. (Guilford Press, New York, 2007), pp. 490–515, second edn.
21. D. M. Wegner, The mind's best trick: how we experience conscious will. *Trends in Cognitive Sciences* **7**, 65–69 (2003).
22. M. I. Norton, D. Mochon, D. Ariely, The IKEA effect: When labor leads to love. *Journal of Consumer Psychology* **22**, 453–460 (2012).
23. N. Castelo, M. W. Bos, D. R. Lehmann, Task-dependent algorithm aversion. *Journal of Marketing Research* **56**, 809–825 (2019).
24. A. R. Doshi, O. P. Hauser, Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* **10** (2024).
25. K. Moon, A. Green, K. Kushlev, Homogenizing effect of large language model on human creative ideation. *Working paper* (2025).

Supplementary Information for

The Agency Trade-Off: Generative AI Reshapes Two Dimensions of Human Agency in Opposite Directions

Benjamin Lira Luttges[1,*], Angela Duckworth[2], Lyle Ungar[3]

[1]The Wharton School, University of Pennsylvania. [2]Department of Psychology, University of Pennsylvania. [3]Department of Computer and Information Science, University of Pennsylvania.

[*]Corresponding author: blira@upenn.edu.

S1 All effects, all data collection efforts

Before turning to the individual pilots, Figure S1 summarizes the Solo–vs.–AI Cohen's d on process agency, outcome agency, and meaning across every data collection effort in this project: the Alex Prolific vignette pilot, the six Prolific cartoon waves (S7a–f), and the two main-paper Gallup studies (filled vs. hollow markers, respectively). Two patterns are visible. First, the process-agency loss is robust: every wave returns a negative d, and the magnitude grows with how much of the visible product the AI authors. Second, the outcome-agency and meaning effects are noisier in the Prolific cartoon waves than in the main-paper Gallup study, because those waves measured all three DVs from the same participant and—as Supplement SS6 shows—rating process agency first contaminated the downstream measurements. The main-paper Study 3 breaks that pattern by randomizing each participant to a single DV block.

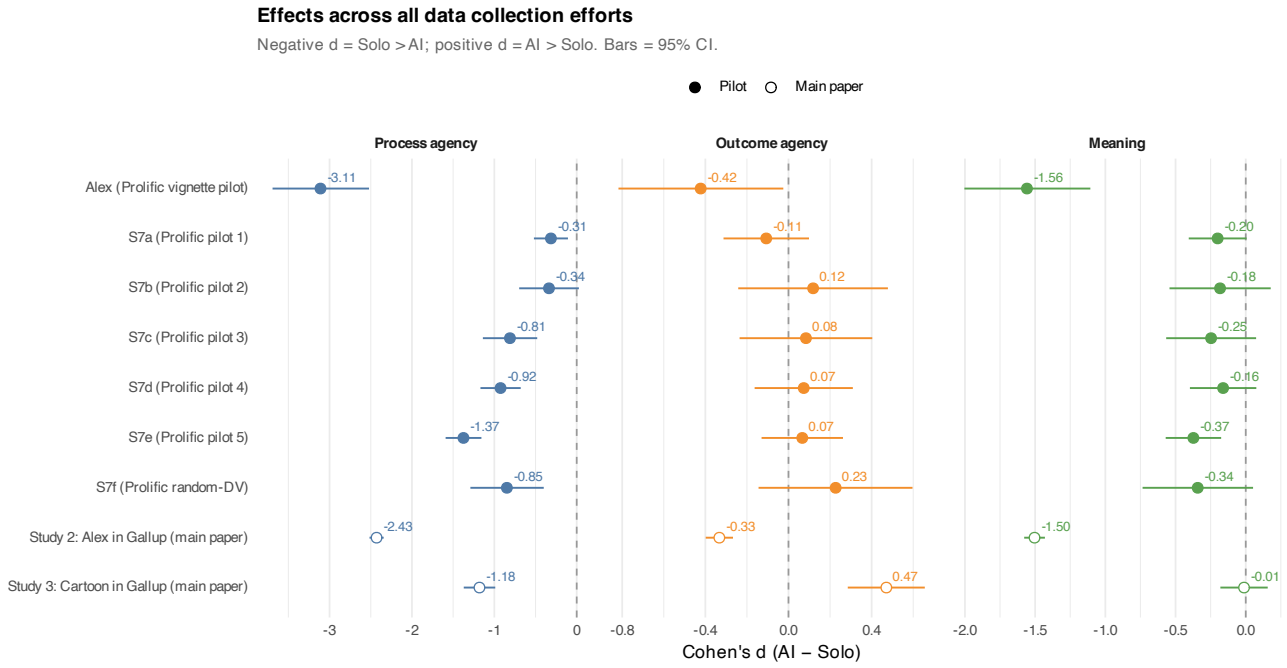

Fig. S1. Solo–vs.–AI Cohen's d on process agency, outcome agency, and meaning across every data collection effort in this project. Filled circles are pilot studies, hollow circles are the main-paper Gallup studies. Bars are 95% CIs. The x-axis range varies by panel.

S2 Sample descriptives and randomization balance

This section reports sample descriptives for each study and randomization-balance checks for the two preregistered experiments (Studies 2 and 3). Study 1 is observational and has no random assignment, so only a descriptives table is reported for it.

Table S1. Study 1 sample descriptives. Participants were U.S. adults recruited through Prolific who had used AI tools for writing in the prior 30 days.

Characteristic	Value
N	399
Age, M (SD)	41.5 (11.9)
Age range	19–81
Gender, female	195 (49%)
Gender, male	198 (50%)
Gender, non-binary	6 (2%)
Employed full-time	260 (65%)
Employed part-time or self-employed	106 (27%)
AI platform: ChatGPT	267 (67%)
AI platform: Gemini	58 (15%)
AI platform: Microsoft Copilot	35 (9%)
AI use frequency: daily	141 (35%)
Work division (% by AI), M (SD)	51.4 (24.4)

Study 1: Factor-analytic sample (N = 399)

Study 2: Gallup vignette sample (N = 3627)

Table S2. Study 2 sample descriptives. Respondents who saw the Alex vignette were Gallup panelists, comprising both Gen Z (18–29) and parent (30+) sample frames.

Characteristic	Value
N (respondents seeing vignette)	3627
Age, M (SD)	33.4 (11.2)
Age range	18–83
Female	792 (61%)
Gen Z sample frame (18–29)	2320 (64%)
Parent sample frame (30+)	1307 (36%)
Race: White	2255 (62%)
Race: Black	63 (2%)
Race: Asian	443 (12%)
Race: Hispanic/Latino	298 (8%)

Table S3. Study 2 randomization balance across the four 2×2 cells. No covariate differed significantly at $p < .05$.

Variable	Solo/Fail	Solo/Succ	AI/Fail	AI/Succ	Test	p
N	904	915	900	908	—	NA
Age, M	33.0	33.5	33.8	33.4	$F(3, 3622) = 0.77$	0.512
% Female	56	62	61	63	$\chi^2(3) = 4.56$	0.207
% Gen Z frame	65	64	62	64	$\chi^2(3) = 1.76$	0.625
% White	65	60	60	64	$\chi^2(3) = 7.38$	0.061

Study 3: Cartoon experiment sample (N = 1533)

Robustness to the sample-frame imbalance. We re-ran the three primary tests of Study 3 controlling for both sample frame (Gen Z self-respondent vs. parent-sample child) and respondent age. The AI vs. Solo coefficient is essentially unchanged on each agency outcome: process agency b = -3.75 ($p < .001$), outcome agency b = 1.36 ($p < .001$), meaning b = -0.06 ($p = .825$). The marginal imbalance therefore does not account for the substantive results reported in the main text.

S3 Study 2: cell means by condition

Cell means and standard deviations for the four 2×2 conditions in Study 2, separately for each of the three dependent variables.

Table S4. Study 3 sample descriptives. Age combines Gen Z self-respondents (18–29) and the parent-sample children's own self-report (12–19, top-coded at 19).

Characteristic	Value
N (cartoon completers)	1533
Age, M (SD)	21.4 (5.8)
Age range	12–29
Female	392 (61%)
Gen Z self-respondents (18–29)	879 (58%)
Parent-sample children (12–19)	639 (42%)
Race: White	935 (62%)
Race: Black	21 (1%)
Race: Asian	176 (12%)
Race: Hispanic/Latino	119 (8%)

Table S5. Study 3 randomization balance: Solo vs. AI. There was a small imbalance in sample-frame membership ($p = .037$); age, gender, and race did not differ at $p < .05$. Robustness check below.

Variable	Solo	AI	Test	p
N	787	746	—	NA
Age, M	21.1	21.6	$t(1514) = -1.72$	0.085
% Female	62	60	$\chi^2(1) = 0.24$	0.623
% Gen Z self-respondent	55	61	$\chi^2(1) = 4.33$	0.037
% White	60	64	$\chi^2(1) = 2.75$	0.097

Table S6. Study 2 cell means (SD).

Cell (Tool / Outcome)	N	Process Agency	Outcome Agency	Meaning
Solo / Failure	904	8.36 (0.16)	3.37 (0.13)	5.65 (0.16)
Solo / Success	915	8.42 (0.16)	8.72 (0.15)	8.51 (0.15)
AI / Failure	900	4.16 (0.15)	2.86 (0.14)	3.18 (0.16)
AI / Success	908	4.09 (0.15)	7.14 (0.16)	4.40 (0.18)

S4 Discriminant validity of process and outcome agency (Study 1)

The Study 1 main text reports that process agency correlates more strongly with meaning than outcome agency does, while outcome agency correlates more strongly with satisfaction. Table S7 formalizes those comparisons with Steiger's Z tests for dependent correlations, and extends them to two additional external criteria.

Table S7. Steiger's Z tests of differential correlations between the two agency dimensions and four external criteria. $\Delta r = r_{\text{Outcome}} - r_{\text{Process}}$; positive values indicate a stronger correlation with outcome agency. $N = 399$.

External variable	r (Process)	r (Outcome)	Δr	Z	p
Outcome credit	0.55	0.72	+0.17	-6.58	$< .001$
Outcome satisfaction	0.54	0.65	+0.11	-3.85	$< .001$
Meaning	0.60	0.54	-0.05	1.85	0.064
AI attitudes (negative)	-0.37	-0.37	+0.00	-0.01	0.993

Outcome credit and outcome satisfaction correlated more strongly with outcome agency than with process agency, both reliably (credit: $\Delta r = +.17$, $Z = -6.58$, $p < .001$; satisfaction: $\Delta r = +.11$, $Z = -3.85$, $p < .001$). Meaning showed the opposite pattern numerically—process agency correlated more strongly with meaning ($r = 0.60$) than outcome agency did ($r = 0.54$)—but the difference fell short of significance ($Z = 1.85$, $p = .064$). Negative attitudes about AI's effect on users (4-item composite, see main text) correlated identically with the two agency dimensions ($r = -0.37$ vs. -0.37; $Z = -0.01$, $p = .993$). Broad concerns about AI's effect on users thus appear to track agency as a whole rather than either facet on its own.

S5 Prolific pilot of the Alex vignette

Before fielding the Study 2 vignette in the Gallup panel, we ran the same 2 × 2 design as a pilot on Prolific (N = 100; cell N = 26–23). The pilot replicated the three main findings reported in the main text:

- **AI use slashed process agency.** d = 3.11 ($p < .001$); Word M = 9.27, ChatGPT M = 3.67.

- **Petition outcome drove outcome agency.** d = 1.83 ($p < .001$).
- **AI penalty on outcome agency was small but reliable.** d = 0.42 ($p = .037$).
- **AI reduced meaning, fully mediated by process agency.** Total AI effect d = 1.56 ($p < .001$); indirect via PA b = -4.83 ($p < .001$), direct b = 0.56 ($p = .520$).

Table S8. Prolific pilot cell means ($M \pm$ SD).

Tool	Outcome	N	Process Agency	Outcome Agency	Meaning
Word	Failure	26	9.07 (1.30)	4.06 (1.90)	5.73 (2.41)
Word	Success	25	9.49 (1.16)	8.90 (0.98)	9.05 (1.05)
ChatGPT	Failure	26	3.38 (1.88)	3.70 (2.48)	2.09 (2.39)
ChatGPT	Success	23	4.01 (2.59)	6.87 (2.70)	4.20 (3.22)

The pilot's small cell sizes and unrepresentative Prolific sample motivated re-fielding the design in the Gallup panel with $\sim$30$\times$ the statistical power. The direction and rank-ordering of effects are identical between the pilot and the main-paper Study 2.

S6 Why we randomized participants to a single DV block in Study 3

Our preregistered Study 3 design randomized each participant to rate *only one* of the three agency outcomes (process, outcome, meaning). This choice was driven by an order-of-measurement artifact we discovered in three earlier Prolific cartoon studies in which each participant rated all three DVs in counterbalanced order. Figure S2 pools those studies and plots the Solo–vs.–AI Cohen's d for each DV broken out by survey position.

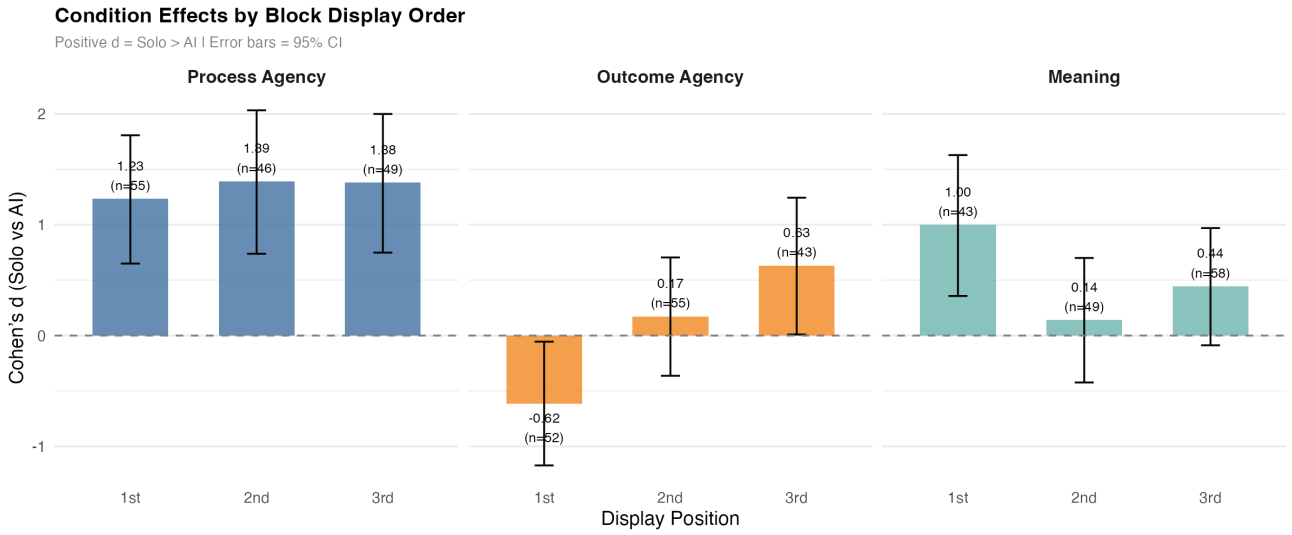

Fig. S2. Solo minus AI Cohen's d on each DV by the position in the survey where the DV was asked (Prolific cartoon studies, pooled). Positive d = Solo > AI. Process agency is roughly position-invariant; outcome agency and meaning flip direction depending on whether process agency was rated first.

Rating process agency first acts as a manipulation in its own right. Once participants have been asked "Did you produce the work?", "Were you responsible for how it was done?", they have been led to attend to the part of their experience where AI most depleted their authorship. That attentional set then contaminates their outcome agency and meaning ratings, flipping the direction of the AI effect for both. We removed this confound in Study 3 by giving each participant only one rating block.

S7 Study 3: full mediation estimates and humor model validation

Parallel mediation. We fit a parallel-mediation SEM in `lavaan` with three mediators (log keystrokes, thumbs-up feedback indicator, max word-Jaccard caption–AI similarity) and a direct AI $\rightarrow y$ path. Coefficients reported in the main text and in Table S9 are fully standardized (`lavaan::standardizedSolution(type = "std.all")`). Bootstrap percentile CIs from 2,000 resamples.

Table S9. Study 3 parallel-mediation indirect effects (standardized).

Path	b	p	DV
AI→keystrokes→PA	-0.18	$< .001$	Process agency
AI→feedback→PA	0.06	$= .015$	Process agency
AI→similarity→PA	-0.66	$< .001$	Process agency
Total indirect on PA	-0.78	$< .001$	Process agency
Direct effect on PA	-0.24	$= .008$	Process agency
AI→keystrokes→OA	-0.00	$= .986$	Outcome agency
AI→feedback→OA	0.32	$< .001$	Outcome agency
AI→similarity→OA	-0.20	$= .076$	Outcome agency
Total indirect on OA	0.11	$= .186$	Outcome agency
Direct effect on OA	0.34	$= .003$	Outcome agency
AI→keystrokes→Mn	0.02	$= .714$	Meaning
AI→feedback→Mn	0.21	$< .001$	Meaning
AI→similarity→Mn	-0.43	$< .001$	Meaning
Total indirect on Meaning	-0.20	$= .008$	Meaning
Direct effect on Meaning	0.19	$= .073$	Meaning

Attrition. Of 1531 adolescents and young adults who submitted a caption, 97.7% completed their assigned rating block. Completion rates were identical in both conditions (97.7% solo, 97.7% AI).

Age moderation (exploratory). We fit $y \sim \text{AI} \times \text{age}_c$ separately for each DV across the 12–29 age range. No interaction reached $p < .05$: process agency $b = +0.043$ ($p = .36$), outcome agency $b = -0.021$ ($p = .65$), meaning $b = -0.069$ ($p = .12$).

Humor scoring model. We fine-tuned GPT-4.1-mini to predict 0–100 percentile-rank humor scores from caption text. On a held-out validation set ($n = 146$), the model achieved Pearson $r = .62$, mean absolute error 17.8 points, and 75% top/bottom-half classification accuracy.

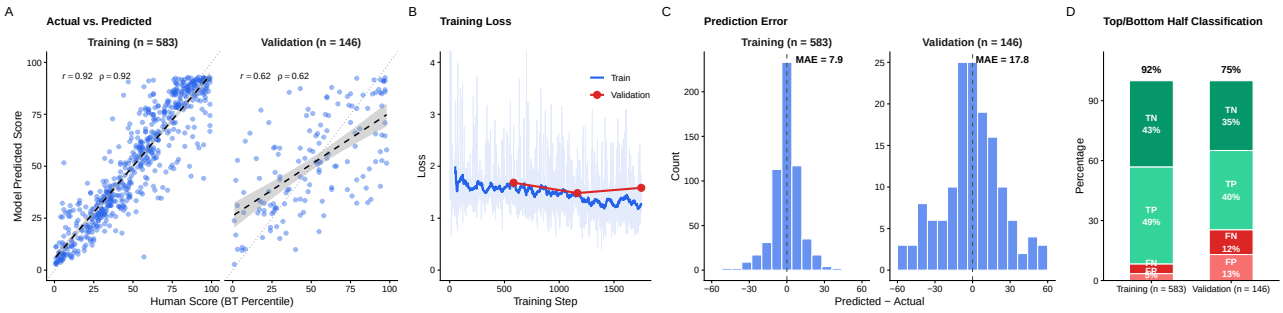

Fig. S3. Humor scoring model validation. (A) Actual vs. predicted scores on training (left) and validation (right) sets. (B) Training loss. (C) Prediction error distribution. (D) Top/bottom half classification accuracy.

S8 Robustness of the Study 3 humor-by-similarity split

Two similarity metrics, same monotone gradient. The main text reports tertiles of caption–AI similarity using word-level Jaccard. Word Jaccard is set-based and ignores order, so a reordered AI sentence still has Jaccard = 1. As a robustness check we recomputed the tertiles using the maximum cosine similarity between the participant's caption embedding and the pool of AI-suggestion embeddings (sentence-transformers `all-MiniLM-L6-v2`, 384 dimensions). The two similarity measures correlate at $r = 0.87$ within the AI condition.

Continuous-cutoff sensitivity. Tertile splits are convenient but arbitrary. To show that the substantive conclusion does not depend on a particular cutoff choice, Figure S5 sweeps the cutoff continuously across the full range of each similarity metric. At each cutoff c, we split AI users into two subsets—those with similarity $\leq c$ and those with similarity $> c$—and plot Cohen's d versus Solo on the pretrained humor score for each subset. Two patterns are visible.

First, the *above-cutoff* curve (dark) is essentially flat in both panels and sits at $d \approx 1.5$ across the entire usable cutoff range: the AI users who adopted any substantial fraction of AI text reliably wrote funnier captions than Solo, and the magnitude is

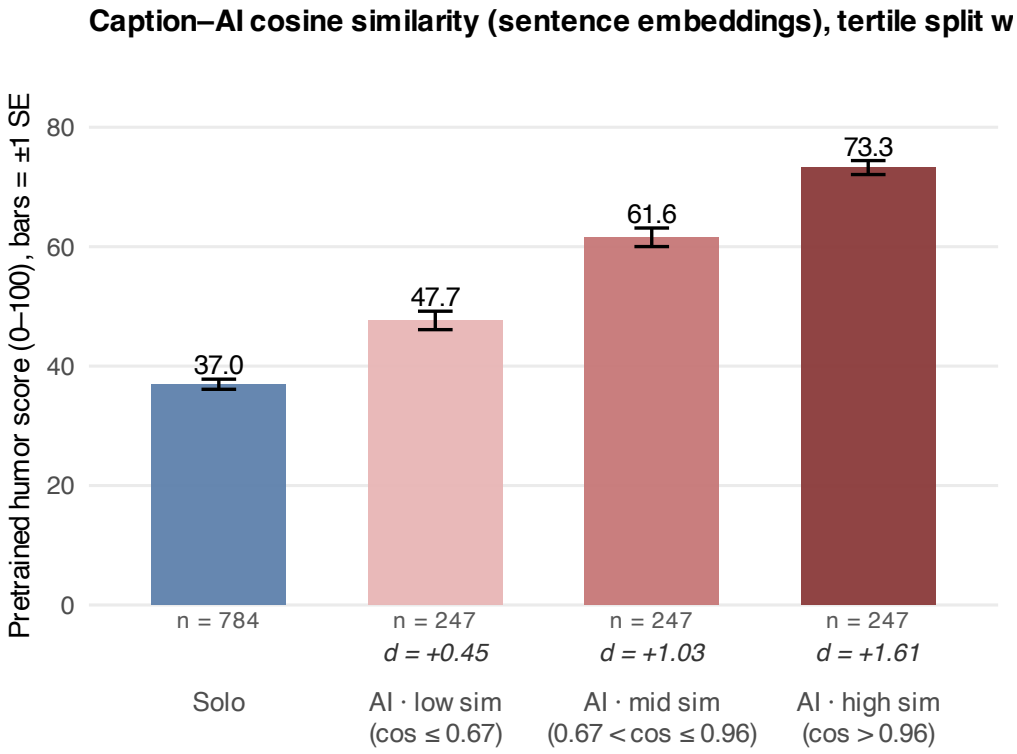

Fig. S4. Pretrained humor score by Solo vs. embedding-cosine tertile of AI participants. Tertile cutoffs (within AI): 0.67 and 0.96. Same monotone gradient as the Jaccard split.

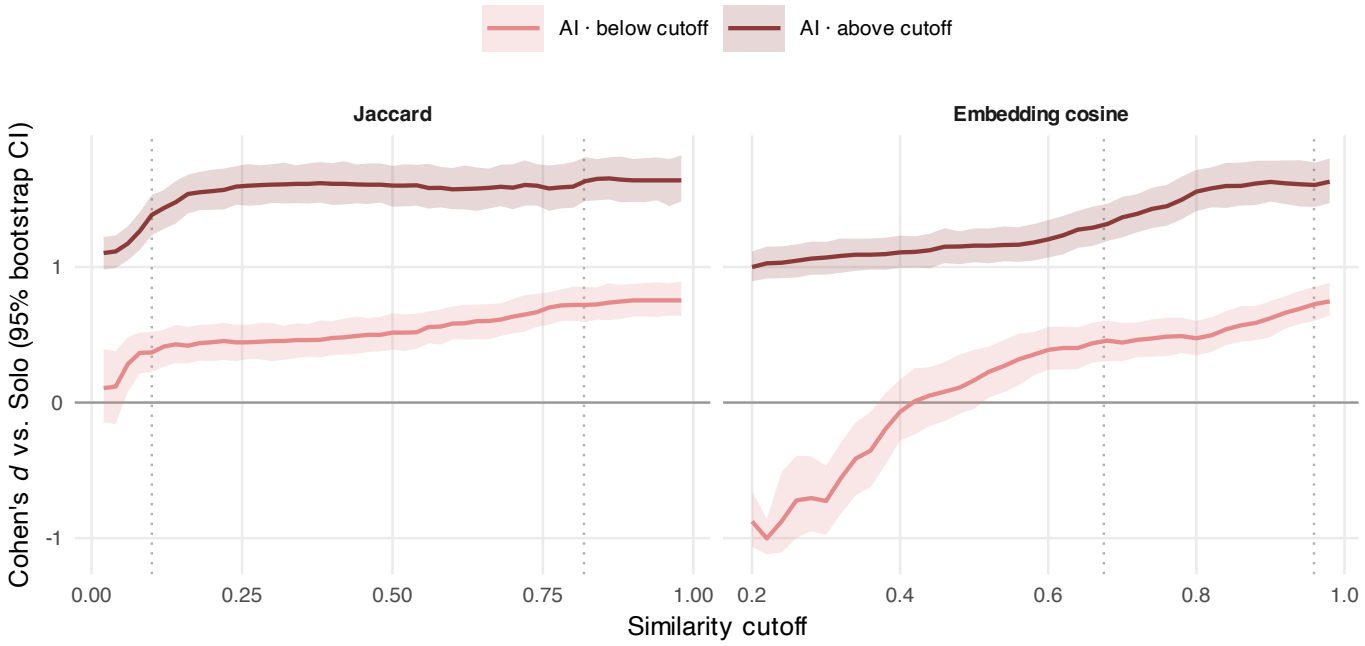

Fig. S5. Continuous cutoff sweep. At every threshold c on the x-axis, we compute Cohen's d (with 500-replicate bootstrap 95% CI) for AI users with similarity $\leq c$ (light line) and similarity $> c$ (dark line) compared to the Solo baseline. Dotted vertical lines mark the 33rd and 67th percentile cutoffs used in the tertile analyses.

insensitive to where one chooses to cut.

Second, the *below-cutoff* curve (light) tells a more nuanced story. Under Jaccard, the below-cutoff subset reliably beats Solo at every threshold ($d > 0$ across the full range), supporting the "every AI subgroup beats Solo" claim in the main text. Under embedding cosine, however, the below-cutoff line dips *negative* at very strict cutoffs (cosine < 0.4; $n \approx 109$ at cutoff 0.5).

The small subset of AI users whose captions had almost no semantic overlap with anything the AI suggested for this cartoon actually scored slightly worse than Solo, plausibly because those captions were off-topic or nonsensical. The Jaccard metric does not isolate this group because "low Jaccard" includes any caption with few overlapping words regardless of topical fit; the embedding metric does. The main-text conclusion (all three Jaccard tertiles beat Solo) therefore holds robustly under any Jaccard cutoff but is restricted to the moderate-to-high range under embedding cosine.

S9 Bradley–Terry calibration of the humor pool

We fit a Bradley–Terry log-strength model $P(\text{win}_{i,j}) = \sigma(s_i - s_j)$ on 14,350 pairwise votes across 729 captions from earlier Prolific cartoon studies. Log-strength parameters were converted to 0–100 percentile ranks against the solo sub-distribution to anchor the fine-tuning target. We split the captions 80/20 into train/validation. Training used GPT-4.1-mini with the OpenAI fine-tuning API.