Adena, Maja; Alabrese, Eleonora; Capozza, Francesco; Leader, Isabelle

Working Paper

AI images, labels and news demand

WZB Discussion Paper, No. SP II 2026-601

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Adena, Maja; Alabrese, Eleonora; Capozza, Francesco; Leader, Isabelle (2026) : AI images, labels and news demand, WZB Discussion Paper, No. SP II 2026-601, Wissenschaftszentrum Berlin für Sozialforschung (WZB), Berlin

This Version is available at:
https://hdl.handle.net/10419/336444

WZB
Berlin Social Science Center

Maja Adena
Eleonora Alabrese
Francesco Capozza
Isabelle Leader

AI Images, Labels and News Demand

Discussion Paper

SP II 2026–601

January 2026

Research Area
Markets and Choice

Research Group
Information, Incentives, Inequality

WZB Berlin Social Science Center
Reichpietschufer 50
10785 Berlin
Germany
www.wzb.eu

Maja Adena, Eleonora Alabrese, Francesco Capozza, Isabelle Leader
AI Images, Labels and News Demand
Discussion Paper SP II 2026–601
Wissenschaftszentrum Berlin für Sozialforschung (2026)

Affiliation of the authors:

Maja Adena
Wissenschaftszentrum Berlin für Sozialforschung

Eleonora Alabrese
University of Bath

Francesco Capozza
University of Barcelona

Isabelle Leader
University of Bath

Abstract

AI Images, Labels and News Demand

by Maja Adena, Eleonora Alabrese, Francesco Capozza, Isabelle Leader

We test whether AI-generated news images affect outlet demand and trust. In a pre-registered experiment with 2,870 UK adults, the same article was paired with a wire-service photo (with/without credit) or a matched AI image (with/without label). Average newsletter demand changes little. Ex-post photo origin recollection is poor, and many believe even the real photo is synthetic. Beliefs drive behavior: thinking the image is AI cuts demand and perceived outlet quality by about 10 p.p., even when the photo is authentic; believing it is real has the opposite effect. Labels modestly reduce penalties but do little to correct mistaken attributions.

Keywords: AI; Demand for News; Trust; Online Experiment.

JEL classification: C81, C93, D83.

Adena: WZB, TU Berlin, CESifo, maja.adena@wzb.eu; Alabrese: University of Bath, CAGE, SAFE, ea2143@bath.ac.uk; Capozza: University of Barcelona, IEB, CESifo, francesco.capozza@ub.edu; Leader: University of Bath ial35@bath.ac.uk. IRB was obtained by the University of Bath with the number 10699-12847. The experiment has been pre-registered on AsPredicted with the number 251218. The authors report no conflict of interest.

1 Introduction

Public trust in news has been low and declining in many countries since 2015, including the UK (Reuters Institute for the Study of Journalism, 2024). When audiences distrust the press, they discount journalistic information and rely more on partisan cues (Ladd, 2010). At the same time, synthetic media is already entering political communication: campaign actors have begun using AI-generated visuals in advertising (Novak, 2023). Major platforms have responded by rolling out AI-content labels, including Meta and TikTok (Nellis, 2024; Paul, 2024), aiming to increase transparency around AI-generated visuals.

Yet generative AI may further influence trust. Photorealistic synthetic faces are increasingly indistinguishable from real ones, and are sometimes judged as more trustworthy, raising concerns for news visuals (Nightingale and Farid, 2022). This led to growing interest in whether disclosure can mitigate misperceptions. Early evidence shows that labels carry meaning for audiences: people penalize headlines labeled as "AI-generated," even when the content is accurate or produced by humans (Altay and Gilardi, 2024), and AI-generated visuals can shift perceptions of political news (Ash et al., 2021; Caprini, 2024; Park et al., 2024). However, the effects of AI on trust remain uncertain, with ongoing debates about how to signal AI content, how audiences interpret these labels, and whether they increase or erode trust (Wittenberg et al., 2025).

While news outlets have long used image post-production without disclosing the underlying technology, generative AI is increasingly used to edit or even create entire visuals.[1] Large-scale audits likewise show that AI systems already support routine news production, often without disclosure (Matich et al., 2025; Russell et al., 2025). Audiences are aware of this shift: many suspect AI involvement yet remain unsure where it occurs, and they express more discomfort with visible generative use than with behind-the-scenes applications (Fletcher and Nielsen, 2024; Reuters Institute for the Study of Journalism, 2024). In this context, an AI-generated image — especially when labelled — may invite inferences on automation, cost-cutting, and reduced editorial oversight, carrying reputational costs that traditional editing does not.

We study whether AI-generated images used by news outlets, and their transparent reporting through labels, affect trust in and engagement with those outlets. In a pre-registered online experiment with 2,870 UK respondents, participants read the same article paired with one of four visuals: a wire-service photo (with or without credit) or a matched AI image (with or without a AI label). The AI and wire-service images are nearly indistinguishable and depict the same scene. Participants know ex-ante that they may later subscribe to the outlet's weekly newsletter. We measure (i) behavioural engagement (newsletter sign-up), and whether it varies with (ii) recognition that the image is AI-generated, and (iii) perceived outlet quality as a potential mechanism. This design isolates the causal effect of AI visuals and labels on outlet perceptions.

We posit three hypotheses. First, relative to the main control (a labelled wire-service

[1] See ADM+S Centre overview for documented cases.

photo), newsletter sign-up will be lower in the AI-image conditions (because attribution signals undermine trust). Second, the unlabelled AI image should have a weaker effect on news demand, since more participants might not detect its synthetic origin. Third, lower perceived outlet quality should mediate reduced engagement.

We first confirm that the treatment affects AI recognition, though it is generally poor. Even for the authentic wire-service photo, respondents often reported the image as AI-generated: 47.2% in the unlabelled control and 42.9% in the labelled control. Thus, adding a source label only modestly reduces mistaken AI attributions. Recognition increases significantly only when the image is AI-generated and labelled as such. These results highlight substantial baseline confusion: authentic photographs are often mistaken for AI, and labels improve recognition only when the image is synthetic, consistent with concerns about the difficulty of distinguishing real and synthetic visuals (Nightingale and Farid, 2022).[2]

On behaviour, we find a precise average null. Baseline newsletter demand in the unlabelled control is 36.5%, and the other treatment arms fall in a narrow band between 37.0 and 39.3%. Neither replacing the original photo with an AI image nor adding a transparent label meaningfully alters sign-ups.[3] Taken at face value, these results indicate that neither images nor labels are a first-order lever on short-run engagement—directly informing the labelling policies rolled out by major platforms (Nellis, 2024; Paul, 2024) and debates about what users infer from AI disclosures (Altay and Gilardi, 2024; Wittenberg et al., 2025).

Behind this aggregate null, however, behaviour differs sharply depending on whether respondents correctly identified the image as AI. When we split the sample by recognition, the sign of the AI effect reverses. Among respondents who did not detect the source of the image correctly, AI imagery raises newsletter demand by about 10 percentage points. Among those who did recognise the image source correctly to be AI, demand instead falls by roughly the same magnitude. AI boosts engagement when it goes unnoticed, but penalises the outlet when recognised. Labels modestly temper this penalty at the margin, without closing the recognition-based gap.

Taken together, the results suggest that what matters is not whether an image is AI, but whether audiences *believe* it is, and current source disclosures do little to correct mistaken attributions. This pattern is consistent with evidence that AI visuals can shift consumption by altering perceptions of the information environment (Campante et al., 2025) especially in political-news settings (Caprini, 2024; Park et al., 2024), while disclosure effects remain mixed (Altay and Gilardi, 2024; Wittenberg et al., 2025). It also parallels research on algorithm aversion even when algorithmic performance is strong (Dietvorst et al., 2015; Dominguez-Castelo et al., 2019).

From the outlet's perspective, these responses matter because they arise even when AI

[2]The average dwell time on the article page was more than one and a half minutes with no differences by treatment. We therefore find attention reasonably high and balanced.

[3]The coefficients on AI, Label, and their interaction are all statistically indistinguishable from zero.

delivers high-quality visuals. Industry surveys report strong enthusiasm for generative tools to reduce repetitive work and expand visual formats (Beckett and Yaseen, 2023; Diakopoulos et al., 2024). Our results show that audiences impose a reputational penalty once AI use is recognised, even though the underlying quality is unchanged.

Beliefs about the outlet move in tandem with behaviour. Among respondents who did not recognise the image as AI, AI visuals increase perceived outlet quality by about 0.50–0.60 standard deviations on the composite quality index. Among those who did recognise the image as AI, the pattern reverses: perceived quality declines by roughly 0.45–0.55 standard deviations. These results are consistent with models linking credibility beliefs to news demand, and evidence using newsletter sign-ups as a behavioural proxy (Chopra et al., 2024). In other words, audiences punish outlets when they *believe* an image is synthetic — even when the image is real — and current labelling practices do little to prevent this, a mechanism with direct relevance for outlets and platforms navigating AI use and disclosure (Nellis, 2024; Paul, 2024; Reuters Institute for the Study of Journalism, 2024).

To interpret these patterns, we develop a simple Bayesian framework in which readers update beliefs about outlet quality when they think the outlet uses AI. The model clarifies when AI can increase demand (if undetected) and when it reduces demand (if recognised), delivering predictions that match the empirical results. These dynamics are consistent with classic models of markets where quality is hard to observe and reputational signals drive behaviour (Klein and Leffler, 1981; Shapiro, 1983). Media outlets invest in credibility for this reason (Gentzkow and Shapiro, 2006; Gentzkow et al., 2025). In our setting, recognised AI use acts as a negative signal about the outlet's type — lowering perceived accuracy, trustworthiness, and engagement — even though the content itself is unchanged. For outlets, the reputational cost of being perceived as "the kind that uses AI" may outweigh the productivity gains that generative tools promise.

Contribution to the literature. These findings speak directly to how AI reshapes trust and news demand in realistic media environments. Our paper contributes to three strands of research.

Trust in media. We connect to economic models in which trust is endogenous and shapes information acquisition and inference (Gentzkow et al., 2025), and to evidence on how news outlets affect accountability and civic life (Snyder and Strömberg, 2010). We provide causal evidence whether the *use of AI-generated images* and their transparent disclosure affect perceived outlet quality and downstream engagement. This complements recent field evidence on AI, trust, and consumption with major news outlets (Campante et al., 2025). Whereas their work shows that making AI-generated misinformation salient can increase engagement by heightening concern about the information environment, we isolate a distinct mechanism: holding text, topic, and outlet constant, *the visual origin of a single image* can alter trust and behaviour, and penalties arise when audiences *believe* an image is synthetic, even when it is authentic.

News demand. We extend experimental work measuring newsletter demand and the role

of accuracy concerns versus belief-confirmation motives (Chopra et al., 2022, 2024; Faia et al., 2024), as well as research on fact-checking and belief updating (Barrera et al., 2020; Henry et al., 2022). We show that the *visual format* and *source labels* causally shape beliefs about outlet quality and actual engagement. Specifically, audiences reduce outlet demand when they think a photo is synthetic (even when it is real) and increase it when they do not. Labels do little to correct mistaken attributions, likely because many viewers fail to attend to them. This isolates how the perceived visual credibility, rather than the text or topic, drives news demand within the same controlled article.

AI and the news ecosystem. We contribute to work on generative AI, platform responses, and the credibility of AI-generated content (Altay and Gilardi, 2024; Caprini, 2024; Park et al., 2024; Wittenberg et al., 2025). Existing studies typically vary text, misinformation exposure, or platform-level policies. Our design instead isolates *source-level* consequences of embedding an AI image within a single controlled article, holding topic, text, and outlet constant. We also complement recent field evidence (Campante et al., 2025) by showing that penalties are belief-driven and can arise even when the image is authentic, while labels do little to prevent mistaken inference. Finally, we relate to emerging work on AI-enabled news customization and its effects on matching and engagement (Chopra et al., 2025). Together, the experiment offers direct causal evidence on how AI visuals shape trust and outlets demand in a realistic media environment.

In the remainder of the paper Section 2 describes the experiment, Section 3 reports results, Section 4 discusses implications and concludes.

2 Experimental Design and Sample

We run a controlled experiment in which respondents read the same news article, varying only the accompanying image and its source label. We begin by describing the sample and recruitment before outlining the detailed design.

2.1 Sample

Participants were recruited from the UK general population on Prolific. Before random assignment, respondents reported basic demographics (age, gender, education, region, employment, ethnicity, political affiliation, and income), allowing us to implement quotas on age, gender, and political affiliation to approximate the UK adult population. The survey also recorded news-consumption habits and familiarity with AI.

To ensure data quality, respondents completed a CAPTCHA, an attention check, and a short screening question on the use of large language models; anyone failing these checks was excluded,

as pre-registered. After these exclusions, the final sample consists of 2,870 respondents (target $N \approx 3{,}000$). Table B.1 compares observable characteristics with UK population benchmarks: the sample is broadly similar, though somewhat more educated and having a higher income on average.

The study was pre-registered on AsPredicted (n. 251218).[4] Prolific's new built-in speed check feature was enabled to automatically exclude respondents completing the survey unreasonably fast relative to the median, ensuring sufficient time-on-task for all included participants. Consistent with this, the median completion time among the final sample was eight minutes. Respondents were paid £1.5 (over £11 per hour), in line with Prolific guidelines.

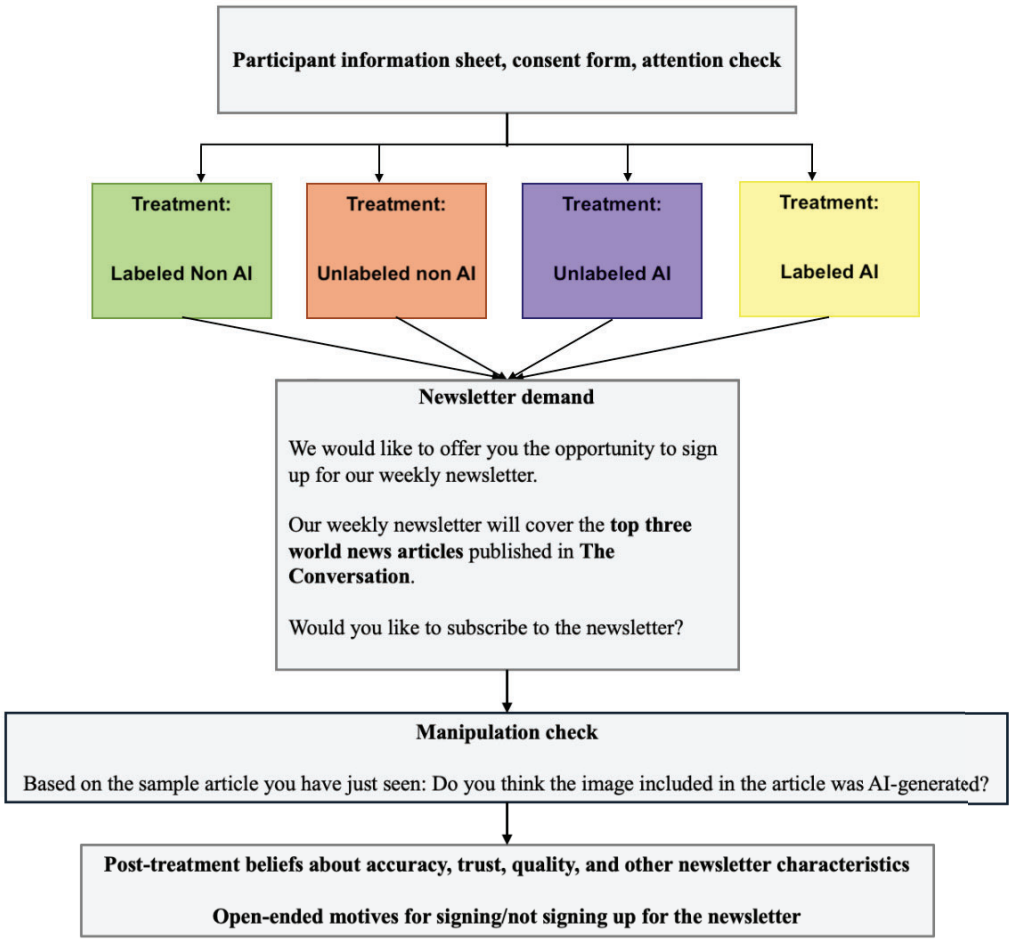
2.2 Experimental Design

After informed consent, standard quality checks, demographics, and media diet questions, each participant reads the same real article from *The Conversation* (UK), a reputable UK academic-led outlet. The article is accompanied by an image. Respondents are randomly assigned to one of four image conditions that vary only in (i) whether the accompanying picture is the original wire-service photograph or a closely matched AI-generated version, and (ii) whether a source label is displayed. Figure 1 provides an overview of the experiment, while Appendix E offers the complete set of instructions.

As an experimental stimulus, we use a real news article from *The Conversation*, a reputable UK academic-led outlet. The source of the news article is made salient at the outset such that subsequent evaluations are clearly attributed to the outlet. Importantly, while reading the article, our participants are aware of the subsequent possibility to subscribe to the weekly newsletter of the same outlet.The article discusses recent protests in California.[5] For a UK sample, the topic is politically relevant yet psychologically distant, encouraging engagement while limiting strong personal identification or sacred-value reactions. Conflict between state and federal authorities also mirrors common themes in contemporary news, with competing narratives and potential for politically motivated misinformation, making it a realistic setting for studying credibility signals.

[4] Pre-analysis plan available at: aspredicted.org/xw49-bcwx.pdf.
[5] Original article available at this link.

Figure 1: Experiment Overview



Treatment manipulation. The experimental manipulation concerns the accompanying image and its source label. In the two control groups, we display the article's original photograph from EPA Images, a major photojournalism agency (Figure 2a). Importantly, the picture carries substantive informational content rather than serving a decorative role, so it plausibly shapes impressions of the outlet. In the two treatment groups, we replace the original photograph with a closely matched AI-generated version created using DALL·E (Figure 2b). Composition, lighting, and subject placement were designed to closely mirror the original so that visual content is comparable aside from image origin and labeling. The full text prompt appears in Appendix D.

Each image appears either with or without a source label. In the **Labelled Control**, the original credit ("*EPA-EFE/Caroline Brehman*") is displayed beneath the photo, as in the real article. In the **Labelled AI** condition, we show: "*AI-generated image (created using OpenAI's DALL·E)*," matching The Conversation's house style and current transparency expectations for AI disclosures (AP, 2023). **Unlabelled** conditions omit any credit or description.

Figure 2: Treatment Manipulation

(a) Original (Control with/without credit)

(b) AI Image (Treatment with/without label)

To avoid deception, respondents are truthfully told that the text is from a real article. Because the possible use of AI images is not mentioned ex ante, participants are debriefed at the end of the survey and provided with full information about synthetic picture and the objective of the study. We judged this omission necessary to preserve naturalistic responses to the article: explicitly warning participants would have primed them to scrutinize the image and undermined the treatment. The content was benign and posed minimal risk. Finally, Table B.2 shows that observable characteristics are balanced across the four experimental conditions.

A natural concern is that we rely on a single picture. However, the intervention does not hinge on idiosyncratic visual features: it varies whether the image is original or AI-generated, and — crucially — how it is described. The AI version is compositionally matched to the original, and the only systematic difference across conditions is the presence or absence of a label. Prior evidence shows that AI labels reduce trust and engagement even when the underlying content is accurate or human-made (Altay and Gilardi, 2024), suggesting that responses generalize beyond any specific stimulus. Field evidence similarly indicates that making AI-generated content salient can shift consumption by altering perceptions of the information environment, a channel that does not depend on the particular creative asset (Campante et al., 2025). Moreover, humans struggle to distinguish synthetic from real visuals across diverse image sets, and photorealistic synthetic faces can even be judged as more trustworthy than real ones (Nightingale and Farid, 2022). Consistent with this broader literature, our effects arise from whether audiences believe an image is AI — updating on the signal that an outlet uses AI, not on which specific picture is shown.

Demand for news (primary outcome). Our primary behavioural outcome is newsletter subscription (yes/no). Prior experimental work validates newsletter sign-up as a revealed-preference proxy for news demand, showing how sign-ups respond to outlet credibility signals (e.g., fact-

checking or outlet-bias cues). We follow this convention and treat sign-up as a compact, incentive-compatible engagement measure (Chopra et al., 2022, 2024). After reading the article coupled with its assigned image, participants choose whether they would like to receive a weekly newsletter summarising the top three world news stories from *The Conversation*. Respondents who opt in indicate interest in receiving a link via Prolific's messaging system, avoiding the need to collect email addresses. While the newsletter is free to access, subscribing still entails a non-monetary opportunity cost, primarily time and attention that could be spent elsewhere. For ethical and data-protection reasons, no messages are actually sent, and participants are informed of this during the debrief.

Manipulation check and recognition coding. After the primary outcome, we measure whether respondents inferred how the image was created. We ask: "*Based on the sample article you just saw, was the image AI-generated?*" (Yes/No; forced choice, with no "Unsure" option). We then construct a recognition indicator:

$$\textit{OriginCorrect} = \begin{cases} 1 & \text{if answer matches the image's true origin (AI} \rightarrow \text{Yes; Real} \rightarrow \text{No)} \\ 0 & \text{otherwise.} \end{cases}$$

Mechanisms (secondary outcomes). After the subscription decision and manipulation check, participants rate the outlet's expected quality on three items: accuracy, trustworthiness, and overall quality (1–5 Likert scales, higher = better). We report results both item-by-item and using a standardized composite quality index (average of the three z-scored items). The index provides a proxy for perceived credibility, which we use to examine whether changes in engagement operate through belief updating about the outlet. In this index, positive values indicate above-average perceived quality and negative values indicate below-average perceived quality, allowing changes to be compared on a common scale.

Additional measures and demand checks To assess alternative mechanisms and possible experimenter-demand effects, respondents also rated perceived political bias, complexity, entertainment value of the article, and trust in the researchers. We additionally included an open-ended question asking participants to guess the purpose of the study (Table B.9). A large share mentioned AI in general terms (53.5%), but almost none referred to the idea that AI images might affect trust in the outlet or willingness to subscribe. Most answers concerned broad themes such as "fake news," "the rise of AI," or "media bias," rather than the treatment itself. This suggests that while AI was salient as a topic, participants did not infer our key behavioural hypothesis, and demand effects are unlikely to drive the results.

3 Results

Across conditions, AI recognition is poor, and both outlet perceptions and demand behaviour hinge on whether people *believe* the image is AI-generated. Labels only marginally reduce mistaken attributions and do not eliminate the divide between those who think the image is AI and those who do not. We expand on these patterns below.

3.1 Manipulation: Poor AI recognition and weak labeling effects

Figure 3a reports positive responses to the question "*Was the image AI-generated?*" (*Manipulation=1*). With the authentic photograph and no credit, 47.2% of respondents said the image was AI. Adding the wire-service credit lowers this only slightly (to 42.9%). Showing an unlabeled AI image produces a similar response (45.8%). The only clear separation appears when the image is both AI *and* labeled: 51.6% of respondents say it was AI. The corresponding regression.(Table B.3) confirms that the AI × Label interaction is positive and statistically significant ($\approx$0.10, $p < 0.01$). Even then, only about half of respondents identify the AI image as AI.

Figure 3b reports the share of correct responses to the same question (*OriginCorrect=1*), indicating whether a respondent correctly identified the image they personally saw as AI-generated or not. The only condition that clearly improves accuracy is the labeled control arm (57.1%), which is significantly higher than the unlabeled AI arm (45.8%). In the labeled-AI arm, the share of correct responses (51.6%) lies between them and is statistically indistinguishable from the two controls.

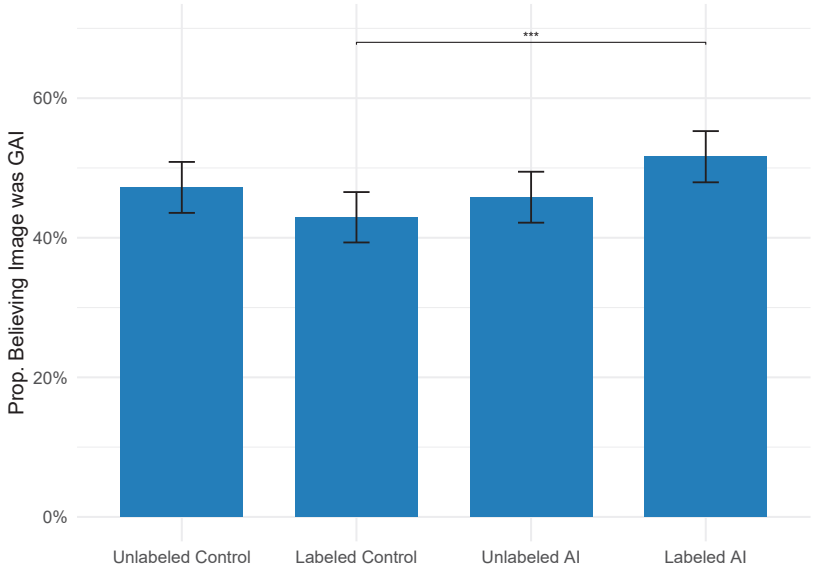
Overall, the manipulation performed as intended: the AI image elicited higher AI attributions once labeled. Yet respondents also frequently attributed AI to the authentic photograph, and labeling had only modest effects on correctness. Importantly, this is not due to inattention. The average dwell time on the article page exceeded 90 seconds in every arm with no systematic differences by treatment (Table B.2), indicating that respondents spent time with the stimulus but still ex-post struggled to recollect whether the image was synthetic or real. This aligns with work showing that photorealistic synthetic images are difficult to distinguish from real photojournalism (Nightingale and Farid, 2022). As the next section shows, these weak recognition responses matter: neither AI images nor labels reduce average engagement, but strong heterogeneities emerge depending on whether respondents believe the image is AI or not.

3.2 Engagement: No average effect of AI or labels

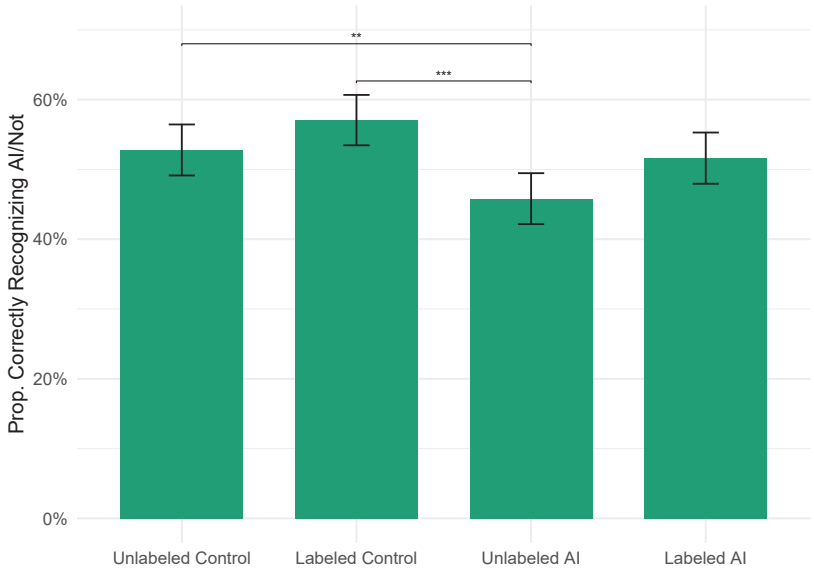
Turning to behavior, Figure 4a plots newsletter demand by treatment arm, and Table B.4 reports the corresponding regressions. In the unlabeled control group, 36.5% of respondents request the newsletter. Assigning an AI image produces almost the same rate (37.0%), adding a label

Figure 3: Manipulation check results.

(a) Share Believing Image was AI Generated (*Manipulation = 1*) by Treatment Arm



(b) Share Identifying AI/Non-AI Correctly (*OriginCorrect = 1*) by Treatment Arm



Notes: Panel (a) shows the share of respondents reporting that the image was AI-generated (*Manipulation = 1*); Panel (b) shows the share who correctly recognized whether the image was AI-generated or not (*OriginCorrect = 1*). Error bars show 95% confidence intervals; stars above brackets denote significance levels (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$).

produces a small and insignificant increase (39.3%), and combining AI with a label yields 37.2%. In the regression, neither AI, nor Label indicators, nor their interaction is statistically different

from zero. The joint test is also clearly insignificant.[6]

With $N = 2{,}870$ and $\alpha = 0.05$ (two-sided), the data exclude *moderate* average effects (larger than 5.5–7.8 p.p.). In similar experiments, Chopra et al. (2022) document very small average shifts (+1.4 p.p. for MSNBC among Democrats), which lie below our minimum detectable effect and are therefore consistent with our null. Chopra et al. (2024) likewise show treatment-induced changes of 5–9 p.p. within politically salient subgroups. Our precision thus rules out effects of the size commonly observed in subgroup analyses, while remaining compatible with near-zero changes in the aggregate sample.

Substantively, holding the article text and outlet constant, swapping a real photo for an AI image does not shift short-run demand. This speaks directly to current policy choices: newsroom labeling practices and platform-level disclosure standards (Nellis, 2024; Paul, 2024) are unlikely to affect behavior *unless* audiences actually notice and internalize the image source. This interpretation is reinforced by the weak recognition patterns documented above, and aligns with mixed evidence on what AI labels communicate and how users act on them (Altay and Gilardi, 2024; Wittenberg et al., 2025).

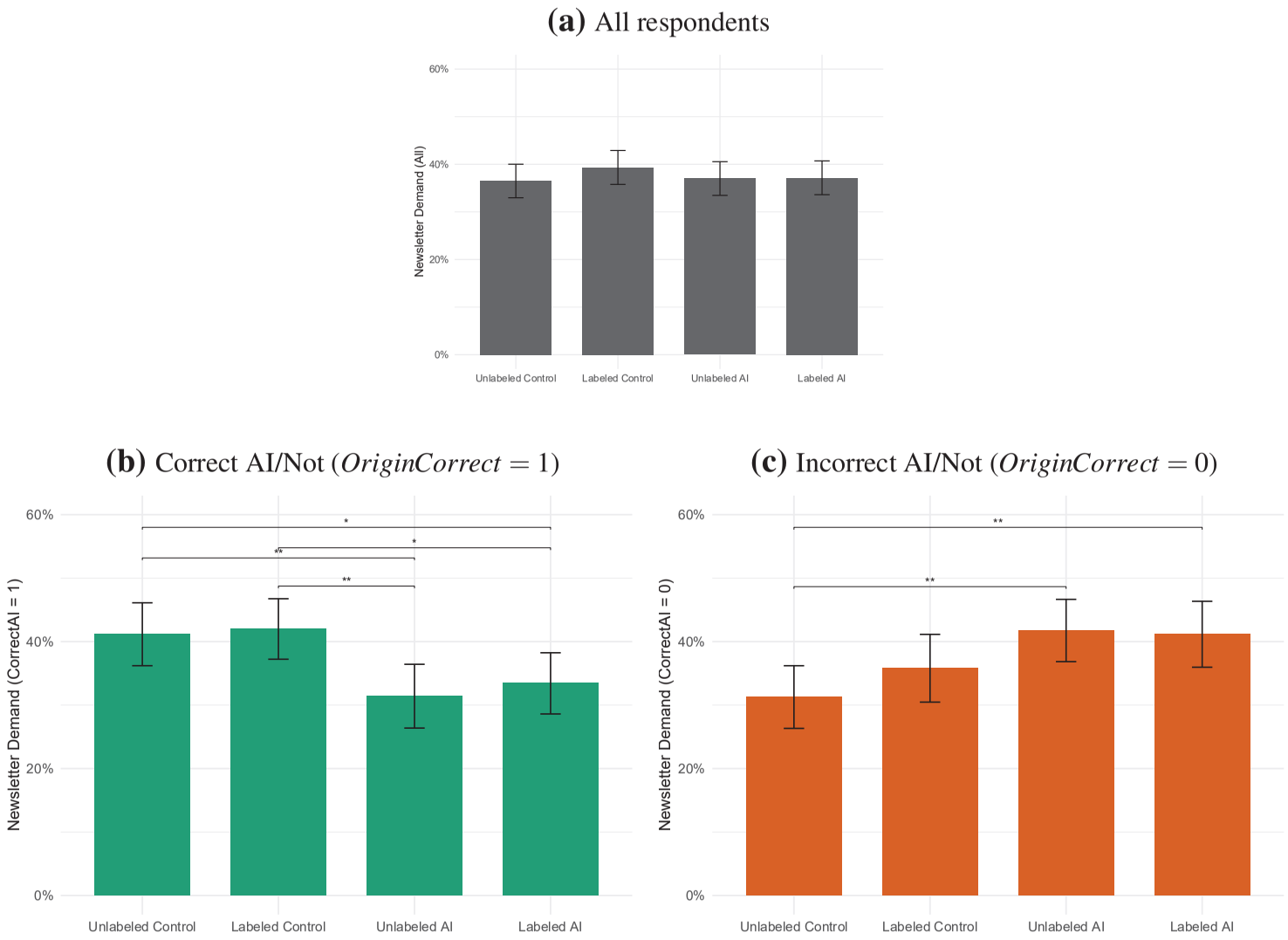
3.3 Heterogeneity: High demand with unnoticed AI, low if recognized

The average null demand effect masks pronounced heterogeneity depending on whether respondents correctly identified the image source (AI vs. not-AI). Figure 4b plots newsletter demand for respondents who correctly identified the image source (*OriginCorrect = 1*), while Figure 4c reports the same outcome for those who did not (*OriginCorrect = 0*). The corresponding regressions appear in Table B.4 (using triple interaction) and Table B.5 (using separate sub-samples).

Among respondents who *did not* correctly identify the image source, assigning an AI image raises newsletter demand by about 10 percentage points (0.081–0.105, $p < 0.01$). Among those who *did* correctly identify the image source, demand instead falls by a similar amount in the case of the AI image (–0.097 to –0.090, $p < 0.01$). These estimates come from the split-sample regressions in Table B.5, and the non-overlapping confidence intervals in both figures (4b and 4c) visually reinforce the sign reversal.

[6]Crucially, while inspecting the article, respondents were aware that they would later have the option to subscribe to the outlet's weekly newsletter. They also spent over 90 seconds on the article page, with no systematic differences across arms (Table B.2), indicating balanced and sufficient attention. As a robustness check (not shown), we re-estimated all outcome effects controlling for reading time; results are unchanged. Likewise, no meaningful heterogeneity emerges above or below the median reading time.

Figure 4: Newsletter Demand by Treatment and AI Recognition Accuracy



Notes: Each panel shows the share of respondents demanding the newsletter by treatment arm. Panel (a) includes all respondents; Panels (b) and (c) split the sample by whether respondents correctly identified whether the image was AI-generated ($OriginCorrect = 1$) or failed to do so ($OriginCorrect = 0$). Error bars represent 95% confidence intervals; stars above brackets denote significance levels (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$).

This pattern is formalized in a simple Bayesian updating framework (Appendix Section A). When respondents correctly identify the image source as AI, they revise their beliefs about outlet quality downward; when they do not, they revise beliefs upward. Demand then moves with these beliefs: higher when the image is not believed to be AI, and lower when it is. Because, in every treatment arm, roughly half of the respondents correctly identify the image and half do not, the positive and negative responses cancel out. As a result, the average effect of each treatment is close to zero, even though the underlying behavioural responses are large.

Labeling does not close the gap between respondents who correctly and incorrectly identified the image source. In the full-sample regression (Table B.4), the three-way AI × Label × OriginCorrect interaction is small and statistically insignificant, and the AI × Label coefficients remain imprecise within each recognition subsample (Table B.5). Substantively, what matters is not whether the image is actually AI, but whether respondents believe it is: AI images sustain engagement when they go unnoticed and penalize the outlet when recognized, and labels do little to correct mistaken attributions. This pattern aligns with evidence that photorealistic synthetic images can appear trustworthy when indistinguishable from real photographs (Nightingale and Farid, 2022) and that AI visuals can shift perceptions in political news contexts (Caprini, 2024; Park et al., 2024).

Another way to see that beliefs drive behavior, and that labels do not meaningfully change outcomes, is in Figures C.1a and C.1b. The figure stratifies newsletter demand solely by whether respondents believed the image was AI-generated (regardless of whether that belief was correct). Two patterns emerge. First, within each belief group — those who believe the image is AI and those who believe it is not — there is no systematic difference across treatment arms: assignment to AI versus control does not move demand conditional on belief. Second, the gap between belief groups is stark. Respondents who believed the image was AI exhibit substantially lower demand than those who believed it was real. This complements Figure 4: treatment assignment generates heterogeneity only insofar as it shifts beliefs, not because AI images or labels have an intrinsic direct effect. In other words, the behavioral differences are belief-driven: what people think they saw matters more than what they were shown. The flat treatment effects within belief groups reinforce that the heterogeneity arises from belief updates rather than from the experimental arm itself.

Two implications follow. *First*, the image source (AI vs. real) does not move average engagement unless users internalise it; recognition is crucial. *Second*, disclosure helps at the margin but is not a remedy, consistent with mixed evidence on labeling regimes (Altay and Gilardi, 2024; Wittenberg et al., 2025) and with platforms' current rollouts (Nellis, 2024; Paul, 2024). For outlets, this creates a trade-off: AI visuals may help when they appear authentic, but may harm trust and engagement when recognised as synthetic, and labels only modestly reduce that risk.

3.4 Mechanism: Credibility and engagement move together

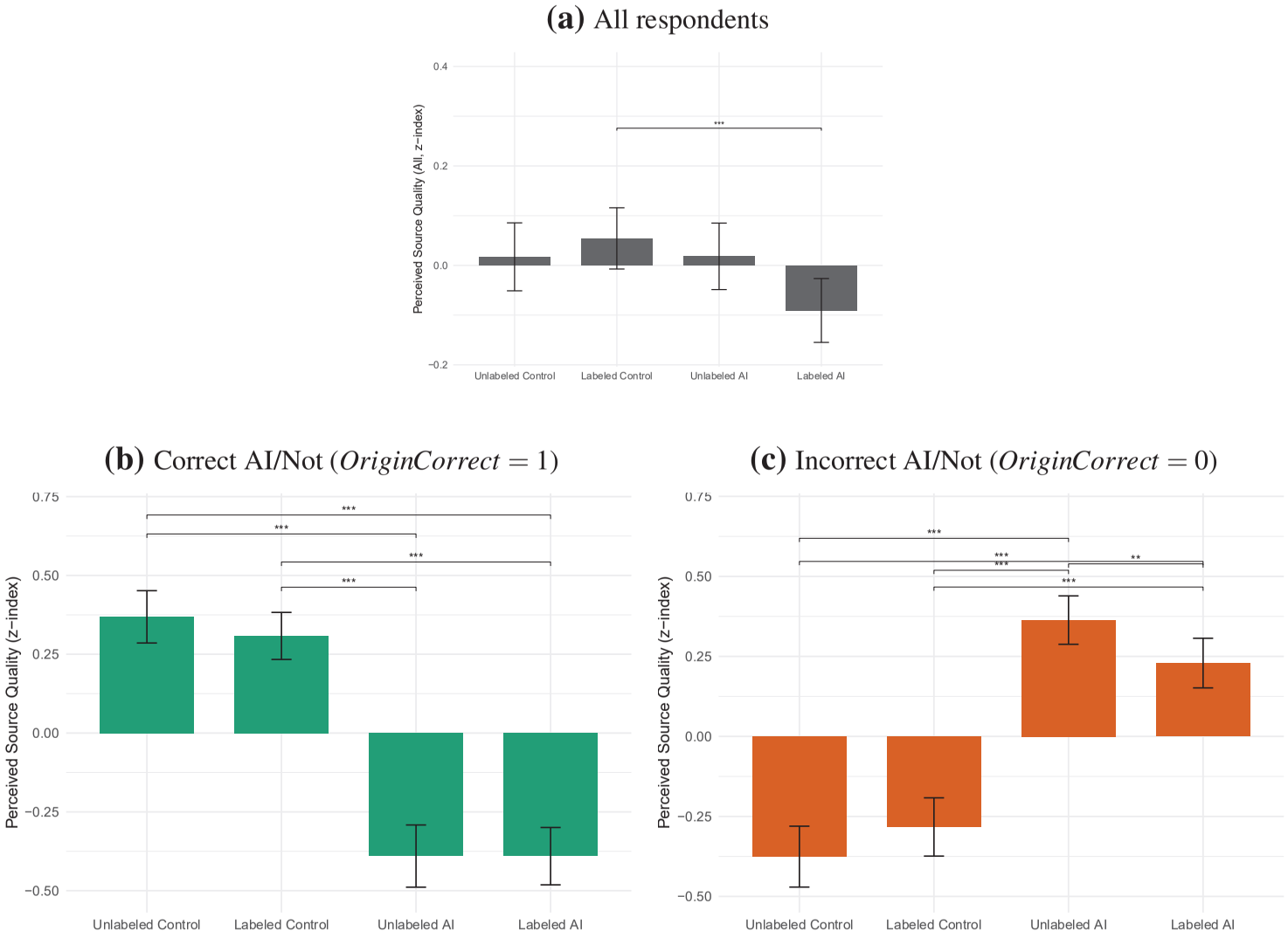
Perceived news source quality mirrors the heterogeneous demand findings. Figure 5a shows that, pooling all respondents, average credibility barely differs across treatment arms: only the labeled AI image is rated significantly lower than the labeled control. This masks sharp differences once we separate respondents by whether they correctly identified the image source.

Table B.6 reports three sets of outcomes: the 1–5 score for *Accuracy*, *Trustworthiness*, and *Overall quality* (columns 7–12); their binary counterparts indicating whether respondents selected one of the two lowest categories (columns 1–6); and a composite credibility index defined as the average of the three standardized items (columns 13–14). Figure 5b and Figure 5c display the same composite graphically for respondents who correctly and incorrectly identified the image source, respectively.

Among respondents who *did not* correctly identify the image source ($OriginCorrect = 0$), the AI image raises perceived credibility (consistent with believing it was a real photograph; see Figure 5c): the composite index increases by roughly 0.6–0.7 standard deviations relative to the control arm (columns 13–14, $p < 0.001$), and each 1–5 item in columns 7–12 shifts upward by a similar margin. Among respondents who *did* recognise the image as AI ($OriginCorrect = 1$), the pattern reverses (see Figure 5b). The interactions in columns 13–14 are strongly negative (around –1.3 to –1.5 SD, all $p < 0.001$), implying net credibility losses of roughly –0.7 to –0.8 SD. Labeling only slightly softens these penalties: for some outcomes, the AI × Label coefficient is weakly negative; for others the triple interaction is weakly positive; but no specification brings perceived quality back to the non-AI baseline.

Viewed through a belief-updating lens, Figures C.1c and C.1d shows perceived outlet quality separately for respondents who believed the image was AI-generated versus those who believed it was not, irrespective of whether that belief was accurate. As with newsletter demand, treatment assignment does not meaningfully move perceived quality within either belief group: respondents who believe the image is real rate the outlet similarly across all treatment arms, and the same holds for those who believe it is AI. The large gap arises between belief groups, not within them. Believing the image is synthetic sharply lowers perceived accuracy, trustworthiness, and overall quality relative to believing it is real, even when both groups were exposed to the same stimulus.

Figure 5: Perceived Outlet Quality by Treatment and AI Recognition Accuracy



Notes: Each panel shows the average perceived outlet quality by treatment arm. Perceived outlet quality is a composite index averaging standardized responses to three 1–5 scale items: *Accuracy*, *Trustworthiness*, and *Quality*. Panel (a) includes all respondents; Panels (b) and (c) split the sample by whether respondents correctly identified whether the image was AI-generated (*OriginCorrect* $= 1$) or failed to do so (*OriginCorrect* $= 0$). Error bars represent 95% confidence intervals; stars above brackets denote significance levels (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$).

Taken together, Figures 4, 5, and C.1 indicate that participants reward the outlet when they believe the image is authentic and penalise it when they believe it is synthetic — even when they are wrong. This credibility pattern mirrors the recognition-contingent demand effects reported in Section 3.3, supporting the interpretation that AI images influence engagement primarily by shifting beliefs about the news source, rather than through their intrinsic visual properties or the presence of a disclosure.

The joint movement of credibility and sign-ups is consistent with models in which perceived outlet quality drives news demand, and with experimental work using newsletter interest as a behavioral proxy (Chopra et al., 2024). In our setting, the operative mechanism is belief updating about the *news source*, conditional on whether respondents think the image is authentic, whether or not the image origin is fully disclosed.

As a complementary test, we treat reported AI recognition as endogenous and instrument it using the randomized AI × Label interaction (Table B.7). The first stage confirms that the interaction raises the likelihood of reporting the image as AI-generated by about 0.10 ($p<0.01$), while the main effects of AI and Label alone are negligible. In the second stage, the 2SLS estimates for newsletter demand are negative but imprecise, consistent with the small reduced-form effects of AI × Label. By contrast, the 2SLS estimates for perceived outlet quality are negative and statistically significant ($p<0.05$), mirroring the reduced-form patterns: respondents pushed into believing the image was AI rate the outlet as less credible. Taken together, both reduced-form and IV evidence point to the same interpretation: AI images influence engagement primarily through beliefs about source authenticity.[7]

We also test several alternative explanations (Table B.8). AI images do not increase perceived political bias or confusion about the text, and respondents do not disengage because they distrust the researcher. Average trust in the researcher remains above "somewhat trustworthy" in every arm, and while trust moves slightly across conditions, the only systematic declines appear when respondents recognise the image as AI (similar to what is observed for source credibility).[8] AI images are also rated as slightly more entertaining on average, with penalties emerging only when respondents recognise the image as synthetic. These patterns rule out concerns that the null average effect or the heterogeneous responses reflect scepticism about the setup or perceived manipulation.

Two additional pieces of evidence support this interpretation. First, open-ended answers

[7]The Wu–Hausman tests do not reject exogeneity of reported AI recognition for either outcome, implying that the IV and OLS estimates are statistically indistinguishable; the IV results therefore reinforce the reduced-form interpretation.

[8]This pattern does not support a deception-based explanation. If respondents distrusted the *experiment*, trust should fall for participants generally exposed to AI, especially in the unlabeled-AI arm, where concerns about being misled would be strongest. Instead, trust rises among respondents who do *not* detect the AI image and falls only among those who do, mirroring the credibility penalty applied to the news outlet. The most consistent interpretation is inference spillover: recognising an AI image lowers beliefs about the outlet and, by association, trust in the presenter of the article, rather than signalling broad experiment distrust.

about the study's purpose (Table B.9) suggest respondents did not guess the hypothesis we test, reducing concern about experimenter demand. A majority state that the study was about "AI or artificial intelligence" (53.5%) or "fake news/media" (24.5%). Very few mention trust or credibility (2.3%), and almost nobody suggests deception (0.1%). Their examples, e.g.: *"Whether humans can tell whether an article is generated by AI"* or *"Fake news"*, indicate respondents took the task at face value and did not anticipate a credibility-focused hypothesis.

Second, open-ended reasons for requesting (or not requesting) the newsletter (Table B.10) align with a neutral, topic-driven interpretation rather than any suspicion. The modal category is curiosity or novelty (47%), followed by interest in AI/technology (33.5%). Very few mention trust or credibility (1.8%). Respondents write, for example, *"I enjoyed reading the article and it gave me a better insight into... "*. These answers reinforce that participants took the task at face value and were responding to the content, not attempting to second-guess the experiment. A small minority even inferred that if the image was synthetic, the article text might also be AI-generated, for example: *"I believe the article is AI generated due to the sentence 'AI-generated image (created using OpenAI's DALL-E)' under the photo. I assume the whole article, not just the photo, is AI generated."* This reinforces that respondents were updating beliefs about the news outlet and its content, not about the researcher or the study design, and only when they recognised the image as AI.

Taken together, alternative mechanisms, guessed study purpose, and open-ended motivations all point in the same direction: respondents interpreted the task straightforwardly, and the demand and credibility effects are driven by whether they believe the image is authentic, not by confusion, scepticism, or experimenter demand.

4 Conclusion

We study whether AI-generated images attached to a real news article alter trust and engagement when text and outlet are held constant. Three findings stand out. First, on average, neither showing an AI image nor adding a label changes newsletter demand relative to a wire-service photograph. Second, effects hinge on what audiences *believe*: when respondents do not realise the image is synthetic, AI imagery raises engagement; when they do recognise it, engagement falls by a comparable amount. Third, outlet perceptions move in parallel. AI imagery improves perceived quality, accuracy, and trustworthiness among respondents who think the image is real, and depresses all three among those who think it is AI. Labels temper this penalty only marginally. Results reconcile an aggregate behavioural null with offsetting belief updates, consistent with news-demand models in which credibility is a proximate driver of engagement (Chopra et al., 2024).

Our design is intentionally narrow to isolate this mechanism: one reputable outlet, one article

on a politically salient but geographically distant event, compositionally matched imagery, and short-run outcomes in a UK online sample. External validity may vary with outlet reputation, topic polarity, image realism, audience priors, and the salience or design of disclosures. These limitations point to concrete next steps: newsroom or platform field experiments; variation in label design and placement; interventions that make provenance salient ex ante (e.g., prebunking or interactive cues); and measurement of revealed engagement (clicks, unsubscribes) across outlets and topics, including political advertising (Caprini, 2024; Novak, 2023; Park et al., 2024).

For practitioners, the implication is pragmatic: if AI imagery is used to illustrate factual reporting by a trusted outlet, average engagement may look unchanged — but only so long as audiences do not notice the synthetic origin. Once they do, belief-based penalties emerge and disclosure may soften credibility costs only at the margin. More broadly, aligning visual production with audience expectations in an era of synthetic media will require pairing transparency with strategies that sustain trust under uncertainty.

References

Altay, Sacha and Fabrizio Gilardi, "People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation," *PNAS Nexus*, 2024, *3* (10), pgae403.

Ash, Elliott, Ruben Durante, Maria Grebenschikova, and Carlo Schwarz, "Visual representation and stereotypes in news media," 2021.

Barrera, Oscar, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya, "Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics," *Journal of Public Economics*, 2020, *182*, 104123.

Beckett, Charlie and Mira Yaseen, "Generating Change: A global survey of what news organisations are doing with artificial intelligence," 2023.

Campante, Filipe R., Ruben Durante, Felix Hagemeister, and Ananya Sen, "GenAI Misinformation, Trust, and News Consumption: Evidence from a Field Experiment," Working Paper 34100, National Bureau of Economic Research August 2025.

Caprini, Giulia, "Visual Bias," University of Oxford, Department of Economics Discussion Paper Series 1016 2024.

Chopra, Felix, Ingar Haaland, and Christopher Roth, "Do people demand fact-checked news? Evidence from U.S. Democrats," *Journal of Public Economics*, 2022, *205*, 104549.

_ , _ , and _ , "The Demand for News: Accuracy Concerns Versus Belief Confirmation Motives," *The Economic Journal*, 2024, *134* (661), 1806–1834.

_ , Ingar K. Haaland, Fabian Roeben, Christopher Roth, and Vanessa Sticher, "News Customization with AI," Working Paper 12121, CESifo Working Paper Series 2025.

Diakopoulos, Nicholas, Seth C. Lewis, and Tamara Witschge, "Generative AI in the newsroom: Adoption, attitudes, and applications," *Digital Journalism*, 2024. Forthcoming.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err," *Journal of Experimental Psychology: General*, 2015, *144* (1), 114–126.

Dominguez-Castelo, Noëmie, T. Alex Bos, and Donald R. Lehmann, "Task-Dependent Algorithm Aversion," *Journal of Marketing Research*, 2019, *56* (5), 809–825.

Faia, Ester, Andreas Fuster, Vincenzo Pezone, and Basit Zafar, "Biases in information selection and processing: Survey evidence from the pandemic," *Review of Economics and Statistics*, 2024, *106* (3), 829–847.

Fletcher, Richard and Rasmus Kleis Nielsen, "Audience perceptions of generative AI in news production," *Reuters Institute for the Study of Journalism Report*, 2024.

Gentzkow, Matthew and Jesse M. Shapiro, "Media Bias and Reputation," *Journal of Political Economy*, 2006, *114* (2), 280–316.

_ , Michael B. Wong, and Allen T. Zhang, "Ideological Bias and Trust in Information Sources," *American Economic Journal: Microeconomics*, 2025, *17* (2), 162–213.

Henry, Emeric, Ekaterina Zhuravskaya, and Sergei Guriev, "Checking and Sharing "Alternative Facts"," *American Economic Journal: Economic Policy*, 2022, *14* (2), 32–60.

Klein, Benjamin and Keith B. Leffler, "The Role of Market Forces in Assuring Contractual Performance," *Journal of Political Economy*, 1981, *89* (4), 615–641.

Ladd, Jonathan M., "The Role of Media Distrust in Partisan Voting," *Political Behavior*, 2010, *32*, 567–585.

Matich, Phoebe, TJ Thomson, and Ryan J Thomas, "Old threats, new name? Generative AI and visual journalism," *Journalism Practice*, 2025, pp. 1–20.

Nellis, Stephen, "TikTok to label AI-generated content from OpenAI and elsewhere," May 2024. Reuters.

Nightingale, Sophie J. and Hany Farid, "AI-synthesized faces are indistinguishable from real faces and more trustworthy," *Proceedings of the National Academy of Sciences*, 2022, *119* (8), e2120481119.

Novak, Matt, "GOP Releases First Ever AI-Created Attack Ad Against President Biden," April 2023. Forbes.

Park, Sohyun, Someen Park, Jaehoon Kim, and Kyungsik Han, "Exploring the Impact of AI-Generated Images on Political News Perception and Understanding," in "Proceedings entry on ACM Digital Library" 2024, pp. 565–571.

Paul, Katie, "Meta to start labeling AI-generated images from companies like OpenAI, Google," February 2024. Reuters.

Reuters Institute for the Study of Journalism, "Digital News Report 2024," 2024.

Russell, Jenna, Marzena Karpinska, Destiny Akinode, Katherine Thai, Bradley Emi, Max Spero, and Mohit Iyyer, "AI use in American newspapers is widespread, uneven, and rarely disclosed," *arXiv preprint arXiv:2510.18774*, 2025.

Shapiro, Carl, "Premiums for High Quality Products as Returns to Reputations," *Quarterly Journal of Economics*, 1983, *98* (4), 659–680.

Snyder, James M. and David Strömberg, "Press Coverage and Political Accountability," *Journal of Political Economy*, 2010, *118* (2), 355–408.

Wittenberg, Chloe, Ziv Epstein, Gabrielle Péloquin-Skulski, Adam J. Berinsky, and David G. Rand, "Labeling AI-generated media online," *PNAS Nexus*, 2025, *4* (6), pgaf170.

For online publication only:

AI Images, Labels and News Demand

Maja Adena Eleonora Alabrese Francesco Capozza Isabelle Leader

A Theoretical Framework

This section adapts the Bayesian news-demand framework in Chopra et al. (2022) to our setting, where the informative object is not "this newsletter is fact-checked" but "this article seems to use an AI-generated image, possibly with a label." The key departure is that the informative signal is *endogenously noisy*: even an authentic photo is frequently believed to be AI (42.9–47.2% in the two control arms), and labels improve recognition only when the image is actually synthetic (51.6% in the AI+label arm). These patterns are documented in Figure 3a and Table B.3.

The model yields four testable implications that line up with the experimental results: (i) a **recognition penalty**: conditional on thinking the image is AI, perceived outlet quality and newsletter demand fall by about 10 p.p.; (ii) a **hidden-AI bonus**: conditional on *not* thinking the image is AI, AI imagery raises perceived outlet quality and demand by a similar amount (Figure C.1, Table B.5 and B.6); (iii) **weak label effects**: labels move recognition only slightly, so they do not close the the demand differences between respondents who correctly and incorrectly identified the image source (Table B.5); and (iv) an **aggregate null**: because recognition is noisy for both real and AI images, the flows into the "high-demand" and "low-demand" posteriors almost cancel, so average newsletter demand stays in the narrow 36.5–39.3% band reported in Figure 4a.

A.1 Environment

There is a single news outlet (the article from *The Conversation* in the experiment). A representative agent decides whether to subscribe to its newsletter:

$$d \in \{0,1\}, \qquad d = 1 \text{ means "subscribe."} \tag{1}$$

The outlet has an unobserved type

$$\theta \in \{H,L\}, \qquad \Pr(\theta = H) = \pi \in (0,1), \tag{2}$$

where H means "high editorial / source standards" and L means "lower standards / more likely to use or mishandle AI visuals." This is exactly the role played by the unobserved outlet quality in Chopra et al. (2022, Appendix A): it governs the value of future news.

If the agent subscribes, her *instrumental* payoff is

$$u(1 \mid \theta) = \begin{cases} v_H & \text{if } \theta = H, \\ v_L & \text{if } \theta = L, \end{cases} \qquad v_H > v_L \geq 0, \tag{3}$$

and $u(0 \mid \theta) = 0$. Interpret v_H as "future newsletters will be accurate / trustworthy / professionally produced," in line with the credibility index in Figure 5. The agent pays an idiosyncratic cost $c \geq 0$ for subscribing (time, inbox clutter, experimenter distrust). Let c be drawn from a continuous CDF F on $\mathbb{R}_+$, independent of signals.

Before deciding, the agent is exogenously shown one of the four visuals used in the experiment:

$$(i, \ell) \in \{R, A\} \times \{0, 1\}, \tag{4}$$

where $i = R$ is the real wire photo, $i = A$ is the matched AI image, $\ell = 1$ denotes the presence of a label (wire credit for R, AI disclosure for A), and $\ell = 0$ denotes no label.

A.2 Signals and recognition

Crucially, the agent does *not* observe i directly. Instead, after seeing (i, ℓ) she forms a binary recognition judgment

$$r \in \{0, 1\}, \qquad r = 1 \text{ means "this looks AI-generated to me."} \tag{5}$$

In the data, r is elicited by the question "Was the image AI-generated?" (forced yes/no), see, Fig. 3a. We model r as a noisy signal about the outlet's type *conditional on the label*:

$$\Pr(r = 1 \mid \theta = H, \ell) = p_{H\ell}, \qquad \Pr(r = 1 \mid \theta = L, \ell) = p_{L\ell}, \qquad \ell \in \{0, 1\}. \tag{6}$$

Assumption A.2 (Diagnostic of "looks AI")**.** For both label conditions,

$$p_{L\ell} > p_{H\ell}.$$

Intuition: low-standard outlets are more likely to put readers in a situation where the visual *looks* or is *read* as AI (because they actually used AI, used it sloppily, or failed to disclose), so seeing $r = 1$ is relatively more likely under $\theta = L$. This is the exact analogue of "biased outlets are more likely to send biased news" in Chopra et al. (2022).

The AI-label experiment further tells us how labels shift these likelihoods. With the authentic photo, adding the wire credit reduced AI attributions from 47.2% to 42.9%; with the AI image, adding an AI label increased AI attributions from 45.8% to 51.6% Fig. 3a. A parsimonious way to encode this is

$$p_{H1} < p_{H0} \approx p_{L0} < p_{L1}. \tag{7}$$

Intuition: labels help a bit when the image is really AI (they push up p_{L1}), but they barely help when the image is real (they push down p_{H1} only slightly), exactly as in the data.

A.3 Bayesian updating

Given (r, ℓ), the agent updates about the outlet's type using Bayes' rule:

$$\mu(r,\ell) := \Pr(\theta = H \mid r,\ell) = \frac{\pi \Pr(r \mid \theta = H,\ell)}{\pi \Pr(r \mid \theta = H,\ell) + (1-\pi)\Pr(r \mid \theta = L,\ell)}. \tag{8}$$

Plugging in (6) gives

$$\mu(1,\ell) = \frac{\pi p_{H\ell}}{\pi p_{H\ell} + (1-\pi)p_{L\ell}}, \tag{9}$$

$$\mu(0,\ell) = \frac{\pi(1-p_{H\ell})}{\pi(1-p_{H\ell}) + (1-\pi)(1-p_{L\ell})}. \tag{10}$$

Lemma A.1 (Recognition is bad news)**.** *Under Assumption A.2,*

$$\mu(1,\ell) < \pi < \mu(0,\ell) \qquad \textit{for } \ell \in \{0,1\}. \tag{11}$$

Proof. For $r = 1$, Assumption A.2 implies $p_{L\ell} > p_{H\ell}$, so the likelihood ratio

$$\frac{\Pr(r=1 \mid \theta = H,\ell)}{\Pr(r=1 \mid \theta = L,\ell)} = \frac{p_{H\ell}}{p_{L\ell}} < 1,$$

which, by Bayes' rule, pushes the posterior below the prior: $\mu(1,\ell) < \pi$. The argument for $r = 0$ is symmetric, using $1 - p_{L\ell} < 1 - p_{H\ell}$. □

Intuition. This is the core insight. In the experiment, respondents punish the outlet *whenever* they think the image is AI, even if it was the authentic wire photo (Figure 4 and Table B.5). The model says: that is rational if "this looks AI" is something only sloppier outlets would make me experience.

A.4 Demand for the newsletter

After observing (r,ℓ), the agent's expected utility from subscribing is

$$EU(1 \mid r,\ell) = \mu(r,\ell)v_H + (1-\mu(r,\ell))v_L - c, \tag{12}$$

and $EU(0 \mid r,\ell) = 0$. She subscribes iff

$$c \le \underbrace{\mu(r,\ell)(v_H - v_L) + v_L}_{=:\bar{c}(r,\ell)}. \tag{13}$$

Hence the (ex ante) probability of subscription in signal cell (r,ℓ) is

$$D(r,\ell) := \Pr(d = 1 \mid r,\ell) = F\big(\bar{c}(r,\ell)\big). \tag{14}$$

Because $\bar{c}(r,\ell)$ is strictly increasing in $\mu(r,\ell)$ and by Lemma A.1 we have $\mu(0,\ell) > \mu(1,\ell)$, we obtain:

Proposition A.2 (Recognition penalty in demand)**.** *For each label condition $\ell \in \{0,1\}$,*

$$D(0,\ell) > D(1,\ell). \tag{15}$$

Intuition and link to the data. This is the 10 p.p. split described in Section 3.3, (Fig. 4 and Table B.5): among respondents who *did* recognise the image as AI, demand falls by ≈ 0.09 to 0.10; among those who *did not*, demand rises by about the same amount. The model produces this because recognition pushes the posterior below the prior, so fewer people cross the subscription threshold.

A.5 From cells to treatment arms

In the experiment we do not observe (r,ℓ) before randomization; we observe only the arm,

$$t \in \{\text{unlabeled real}, \text{labeled real}, \text{unlabeled AI}, \text{labeled AI}\}. \tag{16}$$

Within arm t, the share of respondents who say $r = 1$ is exactly the *manipulation* rate in Fig. 3a. Denote

$$\alpha_t := \Pr(r = 1 \mid t), \qquad 1 - \alpha_t = \Pr(r = 0 \mid t). \tag{17}$$

Average demand in arm t is therefore

$$\bar{D}_t = \alpha_t D(1,\ell(t)) + (1-\alpha_t)D(0,\ell(t)), \tag{18}$$

where $\ell(t) \in \{0,1\}$ is the label status in arm t.

Now plug in the empirical recognition rates:

$$\begin{aligned}\alpha_{\text{unlabeled real}} &\approx 0.472,\\ \alpha_{\text{labeled real}} &\approx 0.429,\\ \alpha_{\text{unlabeled AI}} &\approx 0.458,\\ \alpha_{\text{labeled AI}} &\approx 0.516,\end{aligned}$$

all from Table B.2. These are all in the 0.43–0.52 window, so every arm is a mixture of a *high-demand* state ($r = 0$) and a *low-demand* state ($r = 1$) of very similar size. Because the distance between $D(0,\ell)$ and $D(1,\ell)$ is roughly the same across arms (Prop. A.2), (18) immediately implies:

Proposition A.3 (Aggregate null across arms)**.** *Suppose (i) $\alpha_t \in [0.4, 0.55]$ for all t (as in the data) and (ii) labels do not radically compress the gap $D(0,\ell) - D(1,\ell)$. Then*

$$\bar{D}_t \in [\underline{D}, \overline{D}] \qquad \text{with } \overline{D} - \underline{D} \text{ small.}$$

In particular, parametrizations that match the 10 p.p. recognition gap in Fig. 4b–4c also match their narrow 36.5–39.3% band in Fig. 4a.

Intuition and link to the data. Every arm puts roughly half the sample into the good posterior (high demand) and half into the bad posterior (low demand), so the net effect is small. This is the exact same logic as in Chopra et al. (2022), where fact-checking raises perceived accuracy but average demand barely moves because different respondents update in opposite directions.

A.6 Labels and the two opposite flows

To see why labels "do little to correct mistaken attributions", condition on the two image types:

(i) Real photo. From (7), $p_{H1} < p_{H0}$, so labeling *real* photos shifts some mass from $r = 1$ to $r = 0$, i.e. from $D(1,1)$ to $D(0,1)$. This *raises* average demand in that arm.

(ii) AI image. From (7), $p_{L1} > p_{L0}$, so labeling *AI* images shifts some mass from $r = 0$ to $r = 1$, i.e. from $D(0,1)$ to $D(1,1)$. This *lowers* average demand in that arm.

Because the baseline misrecognition is high for both images, these two flows are of similar size, so the net effect of labeling in the pooled sample is small — exactly what Table A.4 and the insignificant AI × Label × OriginCorrect interaction show.

Proposition A.4 (Why current labels cannot close the gap)**.** *Suppose labels only modestly reduce $p_{H\ell}$ (real photos are still often believed to be AI) and only modestly increase $p_{L\ell}$ (AI images are still often not believed to be AI). Then, even with labels,*

$$D(0,\ell) - D(1,\ell) > 0$$

remains economically large. Hence disclosure tempers but does not remove the recognition penalty.

Intuition and link to the data. Figure 3a shows exactly this: labeling the real photo nudges recognition down by about 4 p.p.; labeling the AI image nudges it up by about 6 p.p. That is far too small to undo a 10 p.p. demand gap generated by recognition (Fig. Fig. 4b–4c).

B Appendix Tables

Table B.1: Summary Statistics: Experiment vs. UK population

Variable (definition)	**Experiment**	**UK Population**
Gender (share female)	52.3%	51.1%
Age (years)	46.7	40.7 (median)
Ethnicity (share White)	88.3%	83.0%
UK-born (share born in the UK)	89.1%	84.0%
Region (share living in the South East)	14.8%	15.6%
Income (share earning > £40K)	41.7%	25.0%
Employment (share employed)	75.8%	75.1%
Education (share with university degree)	62.3%	33.8%
Party (share previously voting Labour)	36.6%	33.7%

Notes: *Experiment* figures refer to our survey respondents. *UK Population* benchmarks draw on the 2021–2022 UK Census (England, Wales, Scotland, Northern Ireland) where possible. When no harmonised UK-wide statistic was available (e.g., income, education, regional population shares), benchmark values are based on the closest nationally representative source. Gender: female share in UK mid-2022; Age: UK median age (ONS); Ethnicity: White share; UK-born: share born in the UK; Region: population in South East; Income: individuals earning above the 75th percentile (HMRC SPI); Employment: employment rate among ages 16–64; Education: share with Level 4+ (England & Wales); Party: Labour vote share in the July 2024 UK General Election.

Table B.2: Balance across Treatments: Group Means, Global and Pairwise Tests

Variable	Unlabeled Control	Labeled Control	Unlabeled AI	Labeled AI	Global_p	Min_pairwise_p
Gender (share of 'Female')	0.54	0.52	0.51	0.53	0.31	1.00
Age	46.36	46.61	47.60	46.20	0.31	0.51
Ethnicity (share of 'White')	0.87	0.89	0.88	0.89	0.17	1.00
UKborn (share of 'Yes')	0.89	0.89	0.89	0.89	0.80	1.00
Region (share of 'South East')	0.15	0.13	0.15	0.16	0.37	0.90
Income (share of '£20,000 - £29,999')	0.15	0.17	0.18	0.17	0.52	0.81
Employment (share of 'Employed full time')	0.51	0.54	0.52	0.53	0.79	1.00
Education (share of 'University or higher')	0.62	0.62	0.62	0.63	0.78	1.00
Political (share of 'Liberal')	0.40	0.40	0.42	0.37	0.22	0.43
Party (share of 'Labour')	0.37	0.36	0.37	0.37	0.96	1.00
InterestNews (share of 'Interested')	0.45	0.45	0.42	0.40	0.42	0.26
ConsumptionNews (share of 'Every day')	0.56	0.59	0.58	0.56	0.79	1.00
ArticleTime (seconds reading article)	93.64	97.46	100.07	103.76	0.18	0.20
			Post-treatment variables			
Manipulation (share believe AI)	0.47	0.43	0.46	0.52	0.01	0.01
OriginCorrect (share identify AI/Not)	0.53	0.57	0.46	0.52	0.00	0.00

Notes: This table reports the mean of each demographic and attitudinal characteristic by treatment arm. For continuous variables, the reported p-values come from an ANOVA test of equality of means across treatment groups. For categorical variables, p-values correspond to Pearson χ^2 tests of independence. The column *Min pairwise p* shows the smallest p-value from pairwise t- or proportion tests between any two treatment arms.

Table B.3: Manipulation Check

	Manipulation ($AI = Yes$)			
	(1) OLS	(2) OLS	(3) Logit	(4) Logit
Constant	0.471***	7.82×10^{-13}	-0.114	-14.6***
	(0.019)	(8.02×10^{-13})	(0.075)	(1.57×10^{-5})
AI	-0.013	-0.025	-0.054	-0.107
	(0.026)	(0.027)	(0.106)	(0.116)
Label	-0.042	-0.042	-0.170	-0.184
	(0.026)	(0.026)	(0.106)	(0.114)
AI × Label	0.100***	0.109***	0.403***	0.472***
	(0.037)	(0.038)	(0.150)	(0.164)
Pseudo R^2	0.00272	0.05798	0.00284	0.06158
Observations	2,870	2,870	2,870	2,870
Controls		x		x
Mean DV	0.469	0.469	0.469	0.469

Notes: *Manipulation* equals 1 if respondents said the item was AI-generated, 0 otherwise. *AI* equals 1 for treatment arms Unlabeled AI (UAI) and Labeled AI (LAI), 0 for control arms (with or without labels). *Label* equals 1 for labeled control and LAI (label present), 0 otherwise. Controls include demographics, political orientation, and media familiarity. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table B.4: AI and Label Effects on Requesting the Newsletter

	Demand (*Newsletter* = *Yes*)							
	(1) OLS	(2) OLS	(3) Logit	(4) Logit	(5) OLS	(6) OLS	(7) Logit	(8) Logit
Constant	0.364***	6.73×10^{-13}	-0.556***	-14.6***	0.313***	-0.082**	-0.788***	-15.0***
	(0.018)	(6.9×10^{-13})	(0.077)	(1.01×10^{-5})	(0.025)	(0.035)	(0.117)	(0.176)
AI	0.006	-0.0007	0.025	-0.005	0.105***	0.083**	0.455***	0.413**
	(0.025)	(0.025)	(0.110)	(0.122)	(0.036)	(0.034)	(0.156)	(0.172)
Label	0.029	0.029	0.123	0.143	0.045	0.036	0.204	0.191
	(0.026)	(0.025)	(0.109)	(0.121)	(0.037)	(0.036)	(0.167)	(0.183)
AI × Label	-0.027	-0.009	-0.116	-0.049	-0.051	-0.017	-0.228	-0.107
	(0.036)	(0.035)	(0.154)	(0.174)	(0.052)	(0.051)	(0.225)	(0.252)
OriginCorrect					0.098***	0.082**	0.426***	0.402**
					(0.036)	(0.035)	(0.157)	(0.176)
AI × OriginCorrect					-0.201***	-0.171***	-0.874***	-0.861***
					(0.051)	(0.050)	(0.222)	(0.249)
Label × OriginCorrect					-0.036	-0.019	-0.165	-0.103
					(0.051)	(0.050)	(0.221)	(0.245)
AI × Label × OriginCorrect					0.062	0.031	0.282	0.191
					(0.072)	(0.070)	(0.313)	(0.352)
Pseudo R^2	0.00037	0.10309	0.00039	0.10844	0.00615	0.10813	0.00646	0.11368
Observations	2,870	2,870	2,870	2,870	2,870	2,870	2,870	2,870
Controls		x		x		x		x
Mean DV	0.375	0.375	0.375	0.375	0.375	0.375	0.375	0.375

Notes: *Demand* equals 1 if respondents said they wish to receive the Newsletter, 0 otherwise. *AI* equals 1 for treatment arms Unlabeled AI (UAI) and Labeled AI (LAI), 0 for control arms (with or without labels). *Label* equals 1 for labeled control and LAI (label present), 0 otherwise. *OriginCorrect* equals 1 if respondents correctly recognized whether the image was AI-generated or not, 0 otherwise. Controls include demographics, political orientation, and media familiarity. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table B.5: AI and Label Effects on Requesting the Newsletter

	Demand (*Newsletter* = *Yes*)							
	OriginCorrect = 1				*OriginCorrect* = 0			
	(1) OLS	(2) OLS	(3) Logit	(4) Logit	(5) OLS	(6) OLS	(7) Logit	(8) Logit
Constant	0.411***	-3.72×10^{-13}	-0.362***	-14.6***	0.313***	1.16***	-0.788***	16.4***
	(0.025)	(3.91×10^{-13})	(0.104)	(4.55×10^{-6})	(0.025)	(0.203)	(0.117)	(1.14)
AI	-0.097***	-0.090**	-0.420***	-0.456**	0.105***	0.081**	0.455***	0.466**
	(0.036)	(0.037)	(0.158)	(0.193)	(0.036)	(0.035)	(0.156)	(0.190)
Label	0.009	0.013	0.039	0.067	0.045	0.037	0.204	0.231
	(0.035)	(0.035)	(0.145)	(0.173)	(0.037)	(0.037)	(0.167)	(0.202)
AI × Label	0.011	0.010	0.054	0.070	-0.051	-0.009	-0.228	-0.102
	(0.050)	(0.050)	(0.218)	(0.262)	(0.052)	(0.053)	(0.225)	(0.282)
Pseudo R^2	0.00654	0.12896	0.00686	0.13710	0.00571	0.16164	0.00602	0.17029
Observations	1,488	1,488	1,488	1,488	1,382	1,382	1,382	1,382
Controls		x		x		x		x
Mean DV	0.375	0.375	0.375	0.375	0.375	0.375	0.375	0.375

Notes: *Demand* equals 1 if respondents said they wish to receive the Newsletter, 0 otherwise. *AI* equals 1 for treatment arms Unlabeled AI (UAI) and Labeled AI (LAI), 0 for control arms (with or without labels). *Label* equals 1 for labeled control and LAI (label present), 0 otherwise. *OriginCorrect* equals 1 if respondents correctly recognized whether the image was AI-generated or not, 0 otherwise. Controls include demographics, political orientation, and media familiarity. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table B.6: Mechanisms: Effects of AI and Label on Perceived Outlet Quality

	NotAccurate		NotTrustworthy		LowQuality		Accurate		Trustworthy		Quality		z_index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Constant	0.186***	0.143***	0.233***	0.172***	0.198***	0.138***	3.03***	3.15***	2.92***	1.69***	3.06***	4.54***	-0.376***	-0.234
	(0.021)	(0.023)	(0.023)	(0.025)	(0.022)	(0.025)	(0.041)	(0.463)	(0.040)	(0.305)	(0.042)	(0.660)	(0.049)	(0.518)
AI	-0.165***	-0.157***	-0.200***	-0.187***	-0.159***	-0.157***	0.602***	0.536***	0.543***	0.476***	0.526***	0.480***	0.739***	0.660***
	(0.022)	(0.022)	(0.025)	(0.024)	(0.024)	(0.024)	(0.052)	(0.049)	(0.052)	(0.049)	(0.055)	(0.054)	(0.062)	(0.058)
Label	-0.037	-0.041	-0.023	-0.020	-0.040	-0.041	0.068	0.067	0.025	0.004	0.119**	0.096	0.093	0.073
	(0.029)	(0.029)	(0.033)	(0.032)	(0.030)	(0.030)	(0.056)	(0.053)	(0.056)	(0.054)	(0.060)	(0.059)	(0.067)	(0.064)
OriginCorrect	-0.154***	-0.143***	-0.188***	-0.172***	-0.142***	-0.138***	0.585***	0.514***	0.586***	0.520***	0.511***	0.458***	0.744***	0.660***
	(0.023)	(0.023)	(0.025)	(0.025)	(0.025)	(0.025)	(0.054)	(0.051)	(0.054)	(0.052)	(0.056)	(0.055)	(0.064)	(0.061)
AI × Label	0.034	0.042	0.025	0.031	0.044	0.047	-0.206***	-0.199***	-0.121	-0.096	-0.187**	-0.143*	-0.227***	-0.194**
	(0.031)	(0.031)	(0.035)	(0.035)	(0.033)	(0.034)	(0.074)	(0.071)	(0.074)	(0.071)	(0.078)	(0.078)	(0.087)	(0.084)
AI × OriginCorrect	0.329***	0.304***	0.402***	0.372***	0.335***	0.323***	-1.20***	-1.06***	-1.14***	-1.01***	-1.05***	-0.959***	-1.50***	-1.34***
	(0.033)	(0.032)	(0.036)	(0.035)	(0.035)	(0.035)	(0.075)	(0.072)	(0.076)	(0.073)	(0.080)	(0.079)	(0.090)	(0.086)
Label × OriginCorrect	0.025	0.029	0.013	0.008	0.026	0.018	-0.115	-0.117*	-0.093	-0.074	-0.140*	-0.114	-0.153*	-0.135
	(0.031)	(0.032)	(0.035)	(0.035)	(0.034)	(0.034)	(0.074)	(0.071)	(0.074)	(0.071)	(0.078)	(0.077)	(0.088)	(0.084)
AI × Label × OriginCorrect	-0.019	-0.024	-0.019	-0.020	-0.045	-0.041	0.253**	0.221**	0.200*	0.153	0.196*	0.127	0.287**	0.222*
	(0.045)	(0.045)	(0.050)	(0.049)	(0.049)	(0.048)	(0.104)	(0.100)	(0.105)	(0.101)	(0.111)	(0.109)	(0.124)	(0.119)
Adjusted R^2	0.07289	0.10212	0.08450	0.14188	0.06044	0.09112	0.13197	0.23147	0.12181	0.22221	0.09699	0.16363	0.14587	0.25182
Observations	2,870	2,870	2,870	2,870	2,870	2,870	2,869	2,869	2,869	2,869	2,869	2,869	2,869	2,869
Controls		x		x		x		x		x		x		x
Mean DV	0.098	0.098	0.129	0.129	0.118	0.118	3.33	3.33	3.20	3.20	3.33	3.33	0	0

Notes: Each column reports OLS estimates of treatment effects on the perceived outlet quality measures. Dependent variables include *Accuracy*, *Trustworthiness*, and *Quality* (measured on 1–5 Likert scales), their binary counterparts (we create a dummy equal to 1 if respondents selected one of the two lowest categories in the Likert scale, 0 otherwise), and standardized versions of these measures (z-scores) averaged into a composite perceived outlet quality *z_index*. *AI* equals 1 for treatment arms Unlabeled AI (UAI) and Labeled AI (LAI), 0 for control arms (with or without labels). *Label* equals 1 for labeled control and LAI (label present), 0 otherwise. *OriginCorrect* equals 1 if respondents correctly recognized whether the image was AI-generated or not, 0 otherwise. Controls include demographics, political orientation, and media familiarity. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table B.7: AI and Label Effects on Requesting the Newsletter and on Perceived Outlet Quality (IV estimates)

	First stage		Second stage			
	Manipulation		Demand IV		z_index	
AI	-0.013	-0.025	0.002	-0.003	-0.019	-0.026
	(0.026)	(0.027)	(0.023)	(0.020)	(0.042)	(0.037)
Label	-0.042	-0.042	0.017	0.025	-0.025	-0.032
	(0.026)	(0.026)	(0.019)	(0.018)	(0.034)	(0.033)
AI × Label	0.100***	0.109***				
	(0.037)	(0.038)				
Manipulation			-0.274	-0.084	-1.448**	-1.317**
			(0.366)	(0.324)	(0.675)	(0.595)
Weak IV F	7.243	8.424				
Wu–Hausman p			0.6010	0.9856	0.2084	0.1957
Adj. R^2	0.003	0.035	-0.030	0.097	-0.039	0.103
Observations	2870	2870	2870	2870	2870	2870
Controls		x		x		x
Mean DV	0.469	0.469	0.375	0.375	0	0

Notes: Columns (1)–(2) report first stage estimates where *Manipulation* is regressed on *AI*, *Label*, and their interaction. Columns (3)–(4) report 2SLS estimates for *Demand*, an indicator for requesting the newsletter. Columns (5)–(6) report 2SLS estimates for *z_index*, an index of perceived outlet quality (higher = better) that averages standardized responses for *Accuracy*, *Trustworthiness*, and *Quality*, each originally measured on 1–5 Likert scales. In the 2SLS specification, *AI* × *Label* is used as the excluded instrument for *Manipulation*. *Manipulation* equals 1 if the respondent identified the image as AI-generated, 0 otherwise. *AI* equals 1 for Unlabeled AI (UAI) and Labeled AI (LAI), 0 for control arms. *Label* equals 1 for labeled control and LAI, 0 otherwise. *Weak–IV F* reports the robust partial F–statistic for the excluded instrument in the first stage, and *Wu–Hausman p* reports the p–value from a robust Durbin–Wu–Hausman exogeneity test for *Manipulation*. Controls include demographics, political orientation, and media familiarity. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table B.8: Mechanisms (Alternative): Effects of AI and Label on Perceived Bias, Complexity, and Engagement

	Bias		Complex		Researchertrust		Entertaining	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	-0.245***	0.758	2.65***	4.61***	3.13***	3.00***	2.40***	1.76***
	(0.045)	(0.472)	(0.043)	(0.408)	(0.042)	(0.291)	(0.044)	(0.395)
AI	-0.052	-0.070	0.011	-0.009	0.414***	0.351***	0.194***	0.154***
	(0.055)	(0.056)	(0.057)	(0.058)	(0.053)	(0.052)	(0.059)	(0.059)
Label	0.019	-0.017	0.012	0.007	0.070	0.042	0.089	0.086
	(0.065)	(0.066)	(0.060)	(0.061)	(0.058)	(0.058)	(0.062)	(0.063)
OriginCorrect	-0.053	-0.063	0.034	0.035	0.359***	0.303***	0.166***	0.111*
	(0.056)	(0.057)	(0.060)	(0.059)	(0.057)	(0.056)	(0.061)	(0.059)
AI × Label	-0.073	-0.027	0.067	0.090	-0.179**	-0.123	-0.206**	-0.175**
	(0.081)	(0.082)	(0.082)	(0.083)	(0.076)	(0.076)	(0.083)	(0.085)
AI × OriginCorrect	0.097	0.107	-0.158*	-0.133	-0.735***	-0.609***	-0.356***	-0.279***
	(0.079)	(0.079)	(0.083)	(0.082)	(0.080)	(0.078)	(0.086)	(0.084)
Label × OriginCorrect	-0.036	0.0001	0.025	0.034	-0.046	-0.024	-0.093	-0.086
	(0.080)	(0.081)	(0.081)	(0.081)	(0.078)	(0.076)	(0.085)	(0.083)
AI × Label × OriginCorrect	0.056	0.007	-0.117	-0.160	0.221**	0.136	0.207*	0.190
	(0.112)	(0.112)	(0.114)	(0.115)	(0.110)	(0.109)	(0.119)	(0.116)
Adjusted R^2	0.00012	0.02806	0.00674	0.03025	0.04422	0.12409	0.00675	0.08805
Observations	2,869	2,869	2,869	2,869	2,869	2,869	2,869	2,869
Controls		x		x		x		x
Mean DV	-0.286	-0.286	2.649	2.649	3.347	3.347	2.492	2.492

Notes: Each column reports OLS estimates for alternative mechanisms, including *Perceived political bias*, *Complexity*, *Entertainment value*, and *Trust in researcher*. All dependent variables are continuous measures, coded on 1–5 Likert-type scales, except for *Bias*, which ranges from −2 (very left-wing) to +2 (very right-wing). *AI* equals 1 for treatment arms Unlabeled AI (UAI) and Labeled AI (LAI), 0 for control arms (with or without labels). *Label* equals 1 for labeled control and LAI (label present), 0 otherwise. *OriginCorrect* equals 1 if respondents correctly recognized whether the image was AI-generated or not, 0 otherwise. Controls include demographics, political orientation, and media familiarity. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table B.9: Guessed Topic of Experiment

Category	Share	Example
AI / Artificial Intelligence	53.50	Whether humans can tell whether an article is generated by AI
Media / News / Social media	24.50	Fake news
Other / Unclear	14.60	No idea.
Politics / Elections / Opinion	4.30	Political awareness
Trust / Credibility / Ethics	2.30	Trust in governments
Psychology / Perception	0.40	attitudes towards immigration
Economics / Finance / Work	0.10	President Trump's decision to deploy 2000 National Guard workers to...
Marketing / Advertising	0.10	Persuasion
Misinformation / Fake news	0.10	To see who believes fake stories

Notes: Shares are percentages of non-empty responses to *"If you were to guess, what is this study about?"*. Categories are assigned via a simple keyword dictionary; the example quote is the first response observed in that category (truncated for space).

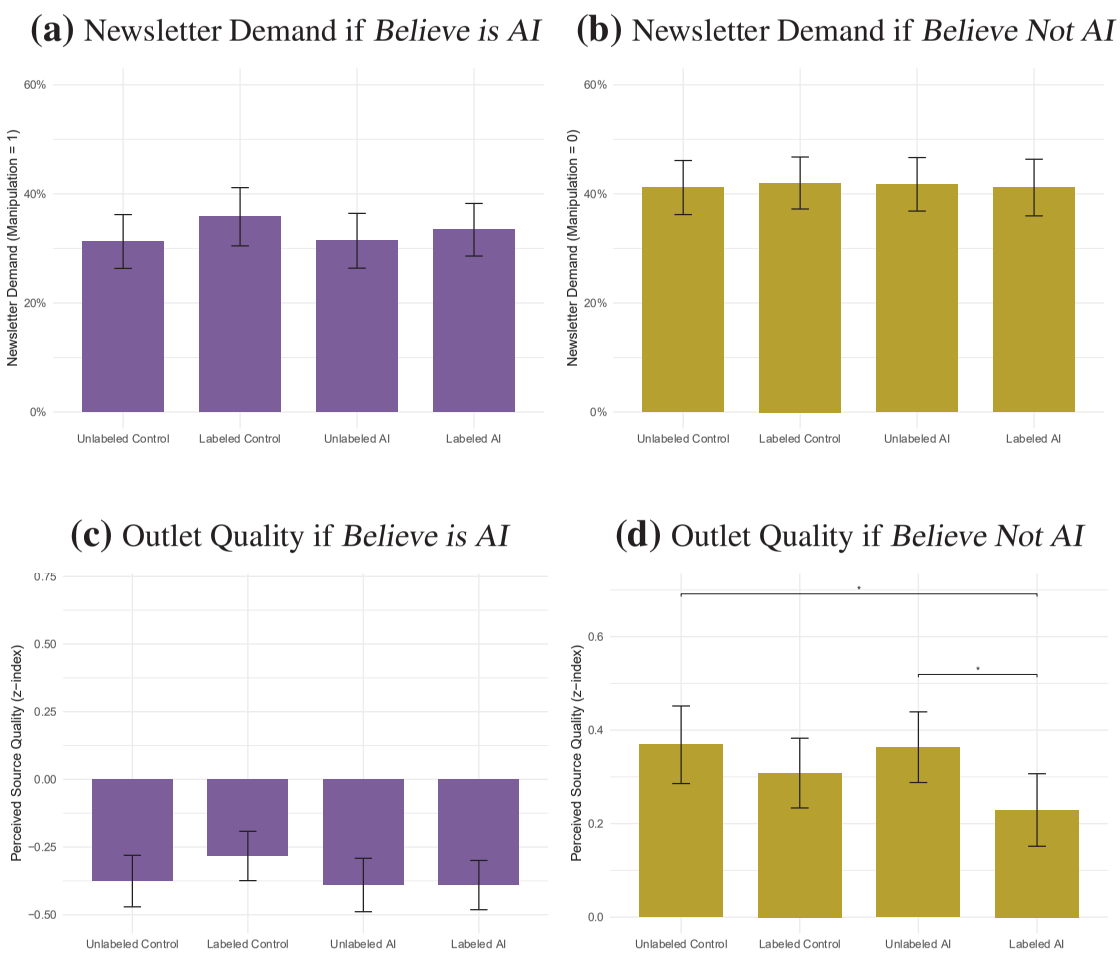
Table B.10: Reasons for Requesting (or Not Requesting) the Newsletter

Category	Share	Example
Curiosity / Novelty	47.00	Not interested, fed up of reading stuff about the US, they would no...
Interest in AI / Technology	33.50	The newsletter seemed wrotten by a human and telling s story about ...
Other / Unclear	13.60	I enjoyed reading the article and it gave me a better insight into ...
Interest in Science / Research	3.40	The story presented seemed interesting and well researched, but I g...
Trust / Credibility	1.80	it seemed like it was on the side of truthful with all the informat...
Skepticism / Disinterest	0.60	I think it was quite a boring read, but I do not enjoy reading a la...
Time / Effort / Relevance	0.20	I found the article too long and drawn out. There was too much info...

Notes: Shares are percentages of non-empty responses to *"Could you please explain your decision regarding the newsletter?"*. Categories are assigned via a keyword dictionary; the example quote is the first response observed in that category (truncated).

C Appendix Figures

Figure C.1: Newsletter Demand and Perceived Outlet Quality by Treatment and AI Beliefs



Notes: Panel (a and b) show the share of respondents demanding the newsletter by treatment arm. Panel (c and d) show the average perceived outlet quality by treatment arm. Perceived outlet quality is a composite index averaging standardized responses to three 1–5 scale items: *Accuracy*, *Trustworthiness*, and *Quality*. Panels (a and c) and (b and d) split the sample by whether respondents believed the image was AI-generated ($Manipulation = 1$) or not ($Manipulation = 0$) respectively. Error bars represent 95% confidence intervals; stars above brackets denote significance levels (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$).

D Instructions to Generate Picture

The full text prompt given to DALL-E to generate the experimental image is as follows:

"please create an AI version of the original image, photorealistic, shot on Canon EOS R5, 85mm lens, f/1.8 aperture, 8K resolution, professional photography, hyper-detailed skin texture, volumetric lighting, HDR".

E Instructions

A full print preview of the Qualtrics survey (including all instructions and blocks) is available at the following link.

Discussion Paper of the Research Area Markets and Choice 2026

Research Group **Information, Incentives, Inequality**

Maja Adena, Eleonora Alabrese, Francesco Capozza, Isabelle Leader SP II 2026-601

AI Images, Labels and News Demand

All discussion papers are downloadable:
http://www.wzb.eu/en/publications/discussion-papers/markets-and-choice