AI Images, Trust and News Demand

Maja Adena Eleonora Alabrese

Francesco Capozza Isabelle Leader Jaime Marques-Pereira

October 20, 2026

Abstract

We study whether AI-generated news images affect demand for and trust in news outlets. Combining descriptive evidence from 353 news organisations across 17 countries with two pre-registered experiments involving over 7,000 UK adults and a survey of 132 international journalists, we document the growing presence of AI imagery in mainstream news (8% of sampled images, likely an underestimate), despite disclosure in fewer than 0.1% of captions. Our experiments vary image origin, source labels, and photorealism, measuring newsletter sign-up and perceived outlet quality. On average, AI imagery and labels do not affect news demand. However, this null masks substantial belief-driven heterogeneity: audiences who recognise an image as AI-generated reduce engagement and downgrade outlet quality by approximately 10 percentage points, while those who do not show the opposite response. Labels do little to correct mistaken attributions. Instrumenting recognition with randomly varied photorealism, we find no effect on newsletter demand but a 0.35-standard-deviation decline in perceived outlet quality. Journalists substantially overestimate the negative effect of AI imagery on audience demand. Our findings suggest that the reputational consequences of generative AI in news depend less on actual image origin than on audiences' beliefs about it, with implications for media credibility and disclosure regulation.

Keywords: AI; Demand for News; Trust; Online Experiment.
JEL: 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, IZA francesco.capozza@ub.edu; **Leader**: University of Bath ial35@bath.ac.uk. **Marques-Pereira**: University of Lancaster j.m.pereira@lancaster.ac.uk. We thank Felix Chopra, Ruben Enikolopov, Marita Freimane, Dorothea Kübler, Ignacio Monzon, Elliot Motte, Mateusz Stalinksi as well as other seminar and workshop participants for their comments. IRB was obtained by the University of Bath with the number 10699-12847. The experiment 1 has been pre-registered on AsPredicted with the number 251218, while the second experiment has been pre-registered on OSF with the number e5ncv. The authors report no conflict of interest. The authors declare no competing interests.

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), and distrustful audiences discount journalistic information in favour of partisan cues (Ladd, 2010). Into this environment has arrived synthetic media: campaign actors now advertise with AI-generated visuals (Novak, 2023), and the major platforms have answered with AI-content labels (Nellis, 2024; Paul, 2024). Disclosure, however, is operationally fragile even where it is policy.[1] Resistance is visible on the producer side too: illustrators have publicly criticised outlets for adopting AI art, arguing that synthetic imagery erodes trust in the reporting it accompanies (Nahas, 2025).[2]

Why this matters beyond the newsroom is clearest in Stiglitz and Ventura-Bolet (2025), who model the information ecosystem as an equilibrium between producers deciding whether to supply truthful content and consumers who differ in their ability to screen quality.[3] When readers cannot tell credible content from the rest, the return to producing it falls and quality erodes. Their analysis takes screening ability as a primitive; we estimate it, and find it in poor shape. Readers classify an authentic wire-service photograph as AI-generated 47% of the time, and the reputational penalty they impose tracks that belief rather than the image's actual origin. Screening failure of this kind taxes outlets that have done nothing wrong and blunts the incentive to invest in credible production on which the equilibrium rests.

Regulation has now moved ahead of the evidence. On 2 August 2026 the transparency chapter of the EU AI Act became applicable, turning AI-content disclosure from a voluntary platform practice into a legal obligation across the Union.[4] European newsrooms are therefore no longer choosing whether to label AI visuals, but how, and at what cost to the outlet and to readers. The evidence base for that choice is thin: what an AI label attached to an otherwise content-equivalent image does to readers' beliefs about the outlet, and to their willingness to consume its news, is essentially unknown.

How common is the practice we study? Section 2 answers this with a large-scale audit: 353 news organisations across 17 countries, drawn from the Reuters Institute *Digital News Report*

[1] In February 2026 Germany's public broadcaster ZDF was found to have aired AI-generated images in its *heute journal* without an on-screen notice. ZDF apologised, removed the segment, and stated that the material should not have been broadcast without clear labelling and editorial context.

[2] Newsrooms nonetheless face the decision with the cost saving in full view: industry surveys report considerable enthusiasm for generative tools as a way to absorb repetitive work and expand visual formats (Beckett and Yaseen, 2023; Diakopoulos et al., 2024), which is precisely why the reputational side of the ledger needs measuring.

[3] AI enters their model through four channels: the efficiency of processing information, the viability of the producer business model, the relative cost of manufacturing misinformation, and consumers' ability to screen quality. Absent adequate regulation, the equilibrium quantity of truthful information falls and polarisation intensifies. The fourth channel is the one an experiment can measure, and the one we target.

[4] Regulation (EU) 2024/1689. Article 50 places the visible-disclosure duty on *deployers*—the category into which a newspaper illustrating an article with a generated image falls—and mandates *that* outlets disclose without prescribing *how*, leaving wording and visual form to a voluntary Code of Practice. The provisions bearing on news production are set out in the second half of this introduction.

and sampled from two independent archives—Common Crawl News and GDELT's Visual Global Knowledge Graph—yielding some 665,000 articles and 1.26 million image occurrences, each scored by AI-image detectors.[5] AI-generated visuals account for 8.0% of weighted image occurrences in CC-News and 2.8% in GDELT; because the classifier favours specificity over sensitivity, these are likely lower bounds. Still a minority, but a growing one: prevalence rises through 2024 in both archives, and the later decline tracks publishers' crawling opt-outs, not falling AI use. Adoption is uneven across countries, from under 1% in Japan to 17.7% in India, and concentrated in soft news: 24% in education and 19% in science and technology against under 2% in crime and disasters. It is not confined to a market tier, being flat in outlets' weekly readership. Disclosure, meanwhile, is close to absent: just 0.09% of 215,638 analysable captions state that the image was AI-generated or name a generative system, and any ambiguous mention of AI raises this only to 0.43%. AI imagery is thus diffusing into routine reporting as an illustrative substitute for stock and wire photography, not as a vehicle for misinformation—and its origin is almost never signalled.

We ask whether AI-generated images used by news outlets, and their disclosure through labels, affect trust in and engagement with those outlets. We run two pre-registered online experiments with UK respondents.[6] Participants read an article paired with one of several visuals: a wire-service photo, with or without credit, or a matched AI image, with or without an AI label. Knowing ex ante that they may later subscribe to the outlet's weekly newsletter, they then decide whether to do so. We measure behavioural engagement (sign-up), whether it varies with recognition that the image is AI-generated, and perceived outlet quality as the candidate mechanism.

We pre-registered three hypotheses: that AI images would depress sign-up relative to a labelled wire photo, that the effect would be weaker for unlabelled AI images because fewer readers detect their origin, and that lower perceived outlet quality would mediate the decline.[7]

The treatment shifts AI recognition, but recognition is poor throughout. Close to half of respondents report the authentic wire-service photograph as AI-generated, and adding a source credit barely reduces the mistake.[8] Authentic photographs are routinely taken for synthetic ones,

[5]The frame spans Europe, the Americas, Asia and Africa, from the UK and US to Brazil, India, Japan, South Africa and Kenya; CC-News runs January 2022–June 2026, GDELT VGKG through November 2024. An image counts as AI-generated when two of Community-Forensics, SAFE and DRCT-2.0 agree, the best-performing rule across three validation datasets; a caption dictionary records disclosure.

[6]$N = 2{,}870$ in Experiment 1 and $N = 4{,}196$ in Experiment 2, after pre-registered exclusions; both fielded on Prolific. Articles are drawn from *The Conversation* and presented in the outlet's own page layout, so respondents evaluate a live article rather than a headline card.

[7]The first and third hypotheses follow from an attribution logic in which observable AI use signals reduced editorial effort; the second anticipates imperfect detection. The results depart from this prior in an informative way. We find no unconditional demand penalty, and detection turns out to be imperfect symmetrically rather than only for AI images, so that the relevant margin is the reader's belief about origin rather than origin itself. The third hypothesis is borne out once recognition is conditioned on. We report all pre-registered tests, including those that do not support our priors, in Section 4.

[8]47.2% in the unlabelled control and 42.9% in the labelled control. Recognition rises significantly only when

and labels improve recognition only when the image is in fact synthetic, consistent with evidence on how hard real and generated visuals have become to tell apart (Nightingale and Farid, 2022).

On behaviour, we find a precise average null. Baseline newsletter demand in the unlabelled control is 36.5%, and every other arm sits within three percentage points of it.[9] Neither replacing the photograph with an AI image nor attaching a transparent label moves sign-ups. Taken at face value, this speaks directly to the labelling policies platforms have rolled out (Nellis, 2024; Paul, 2024), to the duties now binding on European deployers under Article 50, and to the wider debate over what audiences infer from AI disclosure (Altay and Gilardi, 2024; Wittenberg et al., 2025).

Behind the aggregate null, behaviour splits sharply on belief. Readers who mistake an AI image for a photograph sign up about ten percentage points more often than readers who mistake a photograph for an AI image, and among readers who classify correctly the ordering reverses by a similar margin.[10] AI raises engagement when it goes unnoticed and penalises the outlet once it is noticed.

Taken together, 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 is consistent with evidence that AI visuals shift consumption by altering perceptions of the information environment (Campante et al., 2025), especially in political news (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).

Beliefs about the outlet move with behaviour. Perceived outlet quality is roughly half a standard deviation higher among readers who take an AI image to be real than among readers who take a photograph to be synthetic, and the ordering again reverses among readers who classify correctly.[11] Audiences punish outlets when they *believe* an image is synthetic, including when it is not, and current labelling practice does little to prevent the mistake.

To interpret these patterns we develop a simple Bayesian framework in which readers update about outlet quality when they infer AI use. The model clarifies when AI raises demand (undetected) and when it lowers demand (recognised), and its predictions match the data.[12]

the image is both AI-generated and labelled as such. Average dwell time on the article page exceeded ninety seconds with no differences across arms, so the confusion is not an artifact of inattention.

[9] 37.0% for unlabelled AI, 39.3% for the labelled control, 37.2% for labelled AI. The coefficients on AI, Label, and their interaction are individually and jointly indistinguishable from zero.

[10] Labels temper the penalty at the margin but do not close the recognition-based gap. Because recognition is measured post-treatment, these splits are descriptive; Experiment 2 supplies exogenous variation in recognition and recovers the same pattern causally.

[11] Point estimates are $+0.50$ to $+0.60$ SD for the mistaken comparison and -0.45 to -0.55 SD for the correct one. Newsletter sign-up as a behavioural proxy for news demand follows Chopra et al. (2024).

[12] The structure is familiar from markets in which quality is unobservable and reputation carries the signal (Klein and Leffler, 1981; Shapiro, 1983); media outlets invest in credibility for the same reason (Gentzkow and Shapiro, 2006; Gentzkow et al., 2025). Derivations are in the Appendix.

Recognised AI use acts as a negative signal about outlet type, lowering perceived accuracy, trustworthiness and engagement even though the content is unchanged. For an outlet, the cost of being seen as "the kind that uses AI" may exceed the productivity gain the tools deliver.

We then address two natural concerns about Experiment 1: that results may be specific to a single article, and that the belief-based heterogeneity is not causally identified. Experiment 2 preserves the core origin×label design while extending it in two directions.

First, we replace the single article–image pair with six articles spanning distinct topic areas, which lets us assess robustness across contexts.[13] Second, and more importantly, we randomise the photorealism of the AI image: each article appears with its original wire photograph or with one of two matched AI renderings, one highly photorealistic and one deliberately less so, holding the depicted scene and semantic content constant.This shifts the probability that a reader recognises the image as synthetic without altering its origin or its labelling, and so serves as an instrument for recognition rather than requiring us to condition on an endogenous post-treatment belief.

Both findings replicate. Newsletter demand is statistically indistinguishable from control in every arm, and the sign reversal by recognition group—unrecognised AI raising demand, recognised AI reducing it—returns with comparable magnitude. The realism manipulation moves recognition sharply.[14] Instrumenting recognition with it, a one-unit increase in the probability of correct recognition leaves demand unchanged but lowers perceived quality by 0.348 standard deviations.[15]

A natural first reaction to our design is to ask where the harm lies: the AI images we deploy are compositionally matched to the originals, depict the same scene, and contain no misinformation, bias, or manipulation. This is by construction, and it is the source of our identification rather than a limitation of it. Existing evidence that AI imagery erodes trust comes almost entirely from settings in which AI origin is bundled with a content defect—fabricated events, partisan framing, or visibly degraded craft—so that any audience penalty cannot be attributed to provenance as opposed to the deficiency that accompanies it. Holding the depicted scene, the article text, the topic and the outlet fixed, and varying only whether the accompanying image is a wire photograph or a content-equivalent AI rendering, removes these confounds. The question we study is therefore not whether deceptive synthetic media damages trust, which it plainly can, but whether the mere suspicion that a mainstream outlet used an AI image carries a reputational cost with content held constant—which, as the audit shows, is the configuration about which outlets must actually decide.

[13] Business & Economy, Politics & Society, Arts & Culture, Education, and Environment, again sourced from *The Conversation*, paired with twelve AI-generated images in total.

[14] 81.2% of respondents assigned the less realistic image identify it as synthetic, against 44.1% for the photorealistic version, in unlabelled conditions ($p < 0.001$).

[15] 2SLS estimates: $+0.015$ (s.e. 0.050, $p > 0.10$) for sign-up and -0.348 ($p < 0.001$) for perceived quality. Both align in sign and magnitude with the Experiment 1 heterogeneity, supporting a causal reading of the belief channel.

Read this way, our findings are not a non-result. A content-equivalent image still lowers perceived outlet quality by roughly 0.35 standard deviations once recognised as AI, and an authentic photograph is downgraded by a comparable amount whenever it is merely believed to be synthetic. The penalty therefore operates through an inference about outlet type ("the kind of outlet that uses AI maintains weaker editorial standards"), not through anything in the content. Because our images carry no bias or falsity, these effects plausibly represent a lower bound on the reputational cost of recognised AI use.

To see whether practitioners anticipate any of this, we surveyed 132 international journalists.[16] Their beliefs diverge from the data in both directions. They overestimate baseline engagement and expect a large demand penalty from recognised AI that the experiments do not deliver; they almost unanimously support disclosure while underestimating peer support for it and overestimating how widely AI is already used in the profession. These gaps suggest that the reputational risks of AI visual adoption are not yet fully reflected in newsroom decision-making, and that practitioner priors are a useful benchmark against which to read the experimental estimates.

The reduced-form and IV evidence establishes what happens at current levels of AI-detection capability and under the current disclosure regime, but leaves open three questions of practical relevance: how long the aggregate null persists as detection improves, at what audience scale AI adoption stops being financially rational, and what the current noisy-disclosure environment costs relative to the policy a social planner would choose. We address these through three structural exercises, all computed from a single, internally consistent set of Experiment 2 moments to which the journalist survey is also anchored.

The first two are back-of-the-envelope calculations grounded in the model's mixing equation. Parameterising joint improvements in AI detection and real-image authentication by $\lambda \in [0, 1]$, the average demand gap between AI-image and real-image arms is already mildly negative today and reaches -6.2 percentage points under full information.At Experiment 2's scale the gap becomes statistically detectable once roughly 80% of readers identify photorealistic AI images correctly, up from 44% today, and the false-positive rate on authentic photographs falls from 37% to about 13%. Reducing false positives therefore matters as much as improving detection. On the outlet side, AI adoption pays only while the per-image cost saving exceeds the recognition-driven engagement loss, and at standard industry benchmarks there is a critical audience scale—roughly 2,200 effective readers per article—above which AI imagery is unprofitable *today*.

The third exercise recovers the model's structural parameters from the Experiment 2 moments, subjects it to two specification tests not used in estimation, and uses the recovered demand system to compute counterfactual consumer surplus. Full adoption of photorealistic AI images reduces consumer surplus by 2.5% relative to the real-image baseline, with the loss rising to 21% under

[16] Recruited via Prolific; beliefs were elicited against the Experiment 2 baseline sign-up rate, and the elicitation is incentivised. See Berger et al. (2026) for a comparable design surveying journalists.

full information. On optimal disclosure, a planner maximising *any* weighted combination of consumer surplus and outlet profit sets label intensity to zero under the experiment's quality-constant design, since both components fall as labels push readers toward the pessimistic belief state. Separately, the penalty borne by authentic photographs that readers misclassify as synthetic costs roughly 12% of baseline consumer surplus—a loss that image-authentication standards would recover under any welfare metric.

News outlets have long used image post-production without disclosing the technology behind it, but generative AI now edits and even creates entire visuals, and large-scale audits show AI systems already supporting routine news production, often without disclosure (Matich et al., 2025; Russell et al., 2025).[17] Audiences sense the shift without being able to locate it, and report more discomfort with visible generative use than with behind-the-scenes applications (Fletcher and Nielsen, 2024; Reuters Institute for the Study of Journalism, 2024). A visible AI image may therefore invite inferences about automation, cost-cutting and reduced editorial oversight that traditional editing does not.

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

Trust in media. We connect to economic models in which trust is endogenous and shapes information acquisition (Gentzkow et al., 2025), and to evidence on how news outlets affect accountability and civic life (Snyder and Strömberg, 2010). We provide causal evidence on whether the *use of AI-generated images* and their disclosure affect perceived outlet quality and downstream engagement, complementing field evidence on AI, trust and consumption at major outlets (Campante et al., 2025). Whereas that 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* alters trust and behaviour, and penalties arise when audiences *believe* an image is synthetic, even when it is authentic.

Our mechanism also connects to work on the reputational consequences of being seen to use AI. In a labor-market field experiment, Almog (2025) shows that workers under-rely on AI when their use is observable, because visible reliance signals an unfavorable trait to an evaluator even when output accuracy is held constant. We document the audience-side counterpart in a media-trust setting: recognised AI use lowers perceived outlet quality because readers infer that "the kind of outlet that uses AI" maintains weaker editorial standards. The two results trace the same inference on opposite sides of the market, suggesting that the reputational cost of recognised AI use is a general feature of AI adoption rather than an artifact of any single domain.

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

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

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: audiences reduce demand when they think a photo is synthetic, even when it is real, and increase it when they do not. This isolates perceived visual credibility, rather than text or topic, as a driver of 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; Ash et al., 2021; 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 the *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 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 (Chopra et al., 2025).

Our demand-side estimates also connect to fast-growing evidence on how publishers respond to generative AI on the supply side. Zhao and Berman (2026) find that publishers who block GenAI crawlers lose traffic rather than protecting it, and Zhu (2026) shows that the retrenchment is sharply selective by quality and specific to use.[18] Both patterns have an audience-side counterpart here. Readers, like publishers, do not reject AI wholesale: the penalty is conditional on recognition and the average effect on demand is zero. And the outlets with most to lose from being associated with AI in our experiment are those that, in the field, guard their content most aggressively.

The closest demand-side comparison on the text margin is Cho et al. (2026), who exploit the *Wall Street Journal*'s rollout of AI-generated article summaries and find higher recall and confidence in the full article: AI-assisted presentation complements reading rather than substituting for it. Their result and ours are less at odds than they first appear. Disclosed AI that visibly adds something for the reader is rewarded; disclosed AI that replaces something the reader took to be authentic is penalised. What audiences price is not AI involvement as such, but what the outlet used it to do.

Contextualising our estimates in the label-effects literature. Table C.1 places our estimates alongside experimental evidence on label effects from five domains. Two patterns stand out. First, the aggregate null on newsletter sign-up is the norm, not an anomaly: across nutrition labelling (Bleich et al., 2017; Muller et al., 2021), misinformation warning tags (Martel and Rand, 2023; Pennycook and Rand, 2021) and AI-content disclosure (Li and Yang, 2024; Wittenberg

[18] Blocking through `robots.txt` is associated with roughly a 7% reduction in weekly visits. High-factual outlets restrict crawlers at 47% against 8% among low-factual outlets, and higher-quality publishers block training crawlers while leaving search-like access open.

et al., 2025), interventions that shift stated beliefs routinely produce near-zero effects on actual behaviour. Second, our quality-perception effects are large by the standards of that literature. The causal penalty recovered by the photorealism instrument (-0.35 SD) and the conditional effects by recognition group (0.45–0.55 SD) are two to three times the accuracy-perception effects typical of misinformation warning labels (0.10–0.25 SD; Martel and Rand, 2023) and comparable to graphic tobacco warnings (0.20–0.35 SD; Hammond, 2011), suggesting that recognised AI use activates an inference about outlet type rather than merely flagging a potentially inaccurate item.

The belief-driven heterogeneity we document also mirrors the "implied truth effect" (Pennycook et al., 2020): just as tagging some false headlines as disputed leads readers to treat unlabelled content as implicitly validated, the absence of an AI label does not prevent AI attribution here, because 47% of respondents already misidentify an authentic photograph as synthetic. Unlike aspirational labels such as Fair Trade certification, which lift sales and command an 8–10% willingness-to-pay premium (Hainmueller et al., 2015), AI-origin labels are alerting disclosures whose net effect depends on whether audiences update toward the correct belief or, as our IV estimates show, toward a penalty invariant to actual image quality.

Structural estimation and optimal disclosure. We contribute a calibrated demand system for a belief-driven media market, grounding the welfare analysis of AI visual adoption in experimental variation rather than functional-form assumptions. This connects to structural work on media markets in which credibility is a key quality dimension (Gentzkow and Shapiro, 2006), to the literature on credence goods and information asymmetry (Dranove and Jin, 2010), and to mechanism design on optimal disclosure under noisy beliefs (Kamenica and Gentzkow, 2011). The welfare result is no longer counterfactual: Article 50 imposes exactly the mandate our planner would decline to impose in a quality-constant environment, and our estimates quantify what it costs when applied to content-equivalent imagery.

Our welfare analysis also speaks to recent equilibrium evidence on generative AI in creative-goods markets. Goldberg and Lam (2025) find that GenAI crowds out non-GenAI content yet on net raises quality, variety and sales, and that mandatory labelling can forestall a low-quality "market-collapse" equilibrium. Their pro-disclosure conclusion and our finding that mandatory labelling is welfare-reducing are reconciled by the quality assumption: in our quality-constant design labels destroy decision-utility surplus without correcting any payoff-relevant asymmetry, whereas disclosure becomes welfare-improving precisely when AI use is correlated with lower quality, as in their setting (see Appendix B.15). Reimers and Waldfogel (2026) document the same dividing line in book markets, where AI-containing titles receive systematically lower usage and ratings.[19] Where AI use is that strongly correlated with quality, an origin label carries payoff-relevant information.

[19] AI-containing titles already top half of 2025 releases, and author selection accounts for most of the usage differential.

2 Motivational Evidence

Our experiments examine how audiences respond to AI-generated news images and whether explicit disclosure of AI use shapes those responses. These questions are only practically relevant if AI-generated images have become part of routine news production and if disclosure remains imperfect. To establish these conditions, we construct a large-scale observational dataset documenting both the adoption and disclosure of AI-generated images in mainstream online news. We begin by assembling a sampling frame of news outlets across large world economies. For these outlets, we collect news articles from independent sources, extracting article metadata, images, and image captions. We then combine open-source AI-image detectors with caption-based disclosure analysis to estimate both the prevalence of AI-generated images and the frequency with which their use is explicitly acknowledged. Together, these data provide descriptive evidence on the diffusion of AI-generated imagery in mainstream journalism.

Outlet sampling frame. We construct our sampling frame from the Reuters Institute *Digital News Report* (2021–2025). For each of seventeen countries — the United Kingdom, Australia, Germany, France, Italy, United States, Canada, Mexico, Brazil, Argentina, Colombia, Japan, India, South Korea, South Africa, Nigeria and Kenya — we pool all news brands appearing in any annual report over the period. We retain only outlets headquartered in the corresponding country, excluding foreign brands that appear in national readership rankings, and harmonize outlet names across survey waves to account for changes in naming conventions (e.g., broadcast and online editions). Each outlet is then assigned a canonical website and registered internet domain, which serve as the basis for collecting news articles from the sources described below. The resulting sampling frame comprises 353 unique news organizations. Tables A.7, A.8, A.9, A.10, A.11 and A.12 report the complete list of outlets included in the analysis, grouped by country.

Data collection and sampling. We first collect news articles from the Common Crawl News (CC-News) archive, which provides monthly snapshots of publicly accessible news webpages together with their original HTML. Using the canonical internet domains associated with each outlet in our sampling frame, we identify all matching news articles published between January 2022 and June 2026. This procedure yields approximately 82.6 million matched HTML pages across 226 news outlets with CC-News coverage, drawn from the 353 outlets included in the study. We then draw a stratified random sample by country, outlet, and publication year-month, resulting in approximately 407,000 sampled articles. For each sampled article, we parse the HTML, extract all embedded image references together with accompanying metadata (including captions and alternative text), resolve image URLs to their canonical locations, and download the corresponding image files. The sampled articles contain approximately 1.51 million embedded images corresponding to 584,000 unique image URLs, of which approximately 500,000 (about 86%) are successfully retrieved. Throughout the analysis, inverse-probability sampling weights

account for the unequal sampling probabilities induced by the stratified design. Appendix Tables A.13, A.14, A.15, A.16, A.17 and A.18 report descriptive statistics for the CC-News population, while Tables A.19, A.20, A.21, A.22, A.23 and A.24 summarize the corresponding sampled articles and downloaded images.

We complement the CC-News archive with the Global Database of Events, Language and Tone's Visual Global Knowledge Graph (GDELT VGKG), which records images appearing in online news daily, together with machine-generated visual metadata. Using the same outlet domains, we identify all matching articles published between January 2022 and November 2024, yielding approximately 9.8 million matched articles across 256 news outlets with VGKG coverage. To facilitate direct comparisons with the CC-News analysis, we apply the same country–outlet–year–month stratified sampling design, producing a sample of approximately 258,000 articles. Because the VGKG directly records image URLs, no HTML parsing is required; instead, image URLs are extracted directly from the database and downloaded using the same retrieval pipeline employed for the CC-News sample. The sampled articles contain approximately 376,000 image references corresponding to 364,000 unique image URLs, of which approximately 294,000 (80.7%) are successfully downloaded. Appendix Tables A.25, A.26, A.27, A.28, A.29 and A.30 report descriptive statistics for the full VGKG article population, while Tables A.31, A.32, A.33, A.34, A.35 and A.36 summarize the sampled articles and downloaded images.

Finally, we complement these article-level datasets with archived news front pages from the Internet Archive's Wayback Machine. Using the same outlet domains, we identify archived snapshots of outlet front pages between January 2022 and 2026. This procedure yields approximately 22.3 million unique archived snapshots across 320 news outlets with Wayback Machine coverage, drawn from the same 353 outlets included in the study. A snapshot is defined by the combination of its archived timestamp and original URL, allowing us to characterize how frequently each outlet's front page is preserved over time. We use these snapshots to construct a complementary dataset that captures the visual presentation of news at the outlet level, rather than only the images embedded in individual articles. Appendix Tables A.37–A.42 report descriptive statistics for the population of Wayback Machine snapshots, while Tables A.43–A.48 summarize the corresponding random sample.

AI image detection. To identify AI-generated images, we apply four publicly available image-forensic models (Community-Forensics, SAFE, DRCT-2.0, and DRCT-1.4) to every unique downloaded image. These models were developed independently and differ in their underlying architectures and training data, providing complementary approaches to distinguishing synthetic images from conventional photographs. Each detector produces a continuous score indicating the extent to which an image exhibits characteristics associated with AI generation. Given substantial differences in classification outcomes across individual detectors, we construct ensemble classifiers and evaluate their performance using an internal validation sample of

newspaper images and two external datasets that cover different generations of AI-generated imagery.

We evaluate 28 ensemble voting rules and select the one with the highest average Matthews correlation coefficient (MCC) across the three validation datasets. We prioritize MCC because, unlike the F1 score, it explicitly accounts for true negatives, which is particularly important in our setting, where we seek not only to identify AI-generated images but also to avoid incorrectly classifying genuine newspaper images as AI-generated. The selected ensemble classifies an image as AI-generated when at least two of three detectors (Community-Forensics, SAFE, and DRCT-2.0) assign scores of at least 0.5. In the internal validation sample, this rule achieves a sensitivity of 22.64% and a specificity of 98.11%. Its low sensitivity and high specificity indicate that the ensemble is more likely to miss AI-generated images than to incorrectly classify conventional images as AI-generated. Accordingly, the selected ensemble is likely to underestimate the true prevalence of AI-generated imagery in the newspaper corpus. Further details on the detectors, ensemble construction, and validation exercise are provided in Appendix A.

AI disclosure detection. The detection of AI-generated images does not imply that their use was disclosed to readers. To identify explicit disclosure, we analyse figure captions extracted directly from the HTML of sampled CC-News articles. Although we also collect alternative text and other image-related metadata, the disclosure measures reported in this paper are constructed exclusively from figure captions. We identify disclosures using a dictionary of keywords that combines general references to artificial intelligence (e.g., "AI-generated", "generated with AI", and "artificial intelligence") with mentions of widely used generative image models and providers, including ChatGPT, DALL·E, Midjourney, Stable Diffusion, Adobe Firefly, Gemini, Grok, Flux, Ideogram, Imagen, and related systems. The caption dataset and the rule-based classification procedure are described in detail in Appendix A.7.

Prevalence of AI-generated images. Table 1 reports the estimated prevalence of AI-generated images in the CC-News and GDELT VGKG samples. The table first summarizes the number of sampled articles and downloaded images, as well as the number of downloaded images with nonempty figure captions in CC-News, before reporting the number and weighted percentage of image occurrences classified as AI-generated by the selected ensemble.

The CC-News sample comprises 406,853 articles containing 956,911 downloaded image occurrences, of which 230,773 are associated with a nonempty figure caption. We use captioned images for the CC-News prevalence estimates, since a caption indicates that an image belongs to the article itself rather than being, for example, an advertisement. The GDELT VGKG sample includes 258,081 articles containing 303,426 downloaded image occurrences; because VGKG parses images to separate article content from advertisements and other page elements, we use

Table 1: Estimated prevalence of AI-generated images

Statistic	CC-News	GDELT VGKG	Wayback Machine
Sampled articles / snapshots	406,853	258,081	130,901
Downloaded image occurrences	956,906	303,423	2,376,917
Downloaded image occurrences with captions	230,772	–	–
AI-generated image occurrences	12,892	9,161	131,184
AI-generated images (unweighted %)	5.59	3.02	5.52
AI-generated images (weighted %)	7.98	2.75	5.58

Notes: This table reports aggregate statistics for CC-News, GDELT VGKG, and Wayback Machine. CC-News and GDELT VGKG contain sampled news articles, whereas Wayback Machine contains archived newspaper homepage snapshots. For CC-News, the count of downloaded image occurrences with captions is further restricted to images associated with a nonempty figure caption. AI-image prevalence is calculated among these captioned image occurrences for CC-News and among all downloaded image occurrences for GDELT VGKG and Wayback Machine. AI-generated images are images classified as AI-generated by an ensemble of AI detectors, selected through a validation exercise comparing different ensemble rules across internal and external image repositories. Unweighted percentages divide the number of AI-generated image occurrences by the corresponding number of eligible downloaded / captioned image occurrences. Weighted percentages use the same observations and weight each occurrence by the inverse sampling probability of its article or homepage snapshot. Image occurrences include repeated appearances of the same image. For Wayback Machine, the reported number of sampled snapshots refers to sampled homepage snapshots for which image downloads have been attempted so far. Additional homepage snapshots have been sampled, but their images are still being downloaded and scored. Dashes indicate statistics not reported for the corresponding sample.

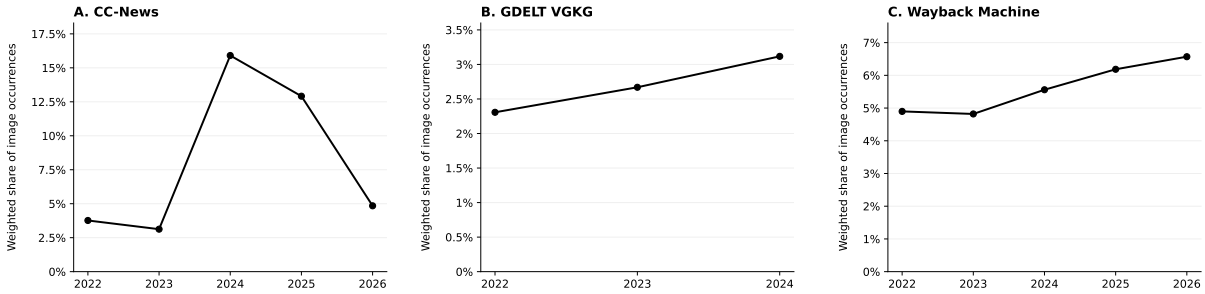
all successfully scored occurrences there.

Applying the selected ensemble, we identify 12,892 image occurrences classified as AI-generated in CC-News and 9,161 in GDELT VGKG. These correspond to 7.98% of eligible image occurrences in CC-News and 2.75% in GDELT VGKG. The classification share is therefore substantially higher in CC-News than in GDELT VGKG, although differences in image collection procedures, sample composition, and coverage periods limit direct comparisons between the two.

Images classified as AI-generated are therefore present in mainstream online news, although they account for a relatively small share of eligible image occurrences. Because the selected ensemble has high specificity but low sensitivity, these shares are likely to understate the actual prevalence of AI-generated imagery.

Figure 1 reports the annual share of image occurrences classified as AI-generated. Both datasets show an increase between 2022 and 2024, although the evolution of the classification share differs considerably across the two samples. In GDELT VGKG, the share increases steadily from approximately 2.3% in 2022 to 3.1% in 2024. In CC-News, it remains relatively stable at approximately 3–4% between 2022 and 2023, before rising sharply to nearly 16% in 2024. These patterns are consistent with an increasing presence of AI-generated imagery in mainstream online news through 2024, although the magnitude and timing of the increase differ across datasets.

Figure 1: Estimated prevalence of AI-generated images over time.



Notes: This figure reports the annual weighted share of downloaded image occurrences classified as AI-generated by the selected ensemble in the CC-News and GDELT VGKG samples. An image is classified as AI-generated when at least two of Community-Forensics, SAFE, and DRCT-2.0 assign it a score of at least 0.50, the rule with the highest mean Matthews correlation coefficient across an internal validation sample of newspaper images and two external validation datasets (Appendix A.3). For CC-News, estimates are restricted to downloaded image occurrences associated with a nonempty figure caption. For GDELT VGKG, estimates use all downloaded image occurrences. In both samples, estimates are restricted to images with complete detector scores. Weighted percentages use the inverse article-sampling probability associated with each image occurrence. Downloaded image occurrences count repeated appearances of the same image across different articles or multiple appearances within the same article.

After 2024, the CC-News classification share declines to approximately 13% in 2025 and 5% in 2026. This decline should, however, be interpreted with caution. Unlike GDELT, whose coverage ends in 2024, the CC-News archive extends through mid-2026, a period during which many large news publishers increasingly restricted automated web crawling policies in response to generative AI (Zhao and Berman, 2025).

This response by news publishers had a severe impact on the coverage of the CC-News dataset, as shown in Appendix Tables A.13–A.18. CC-News coverage declines sharply after 2024, both in the number of outlet domains represented (a decline of 42.2% between 2024 and 2026) and in the volume of HTML pages collected across nearly all countries (a decline of 74.8% over the same period).[20]

These changes are consistent with increasing publisher opt-outs from large-scale web crawling and imply that the composition of the CC-News archive changes substantially over time. Consequently, the post-2024 decline may partly reflect changes in sample coverage and outlet composition rather than a genuine reduction in the use of AI-generated images, and the temporal patterns should be read as descriptive trends within the observed samples.

Disclosure of AI-generated images by news outlets. Appendix Table A.49 summarizes the prevalence of AI-related disclosures in image captions in the CC-News sample. Because disclosure is identified from caption text, the analysis is restricted to figure occurrences with captions that were originally written in English or successfully translated into English before

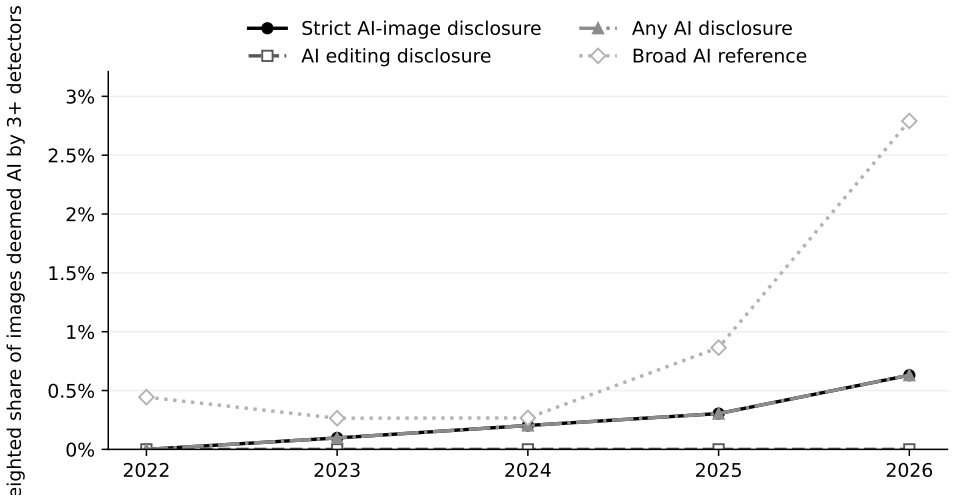
[20]For example, the number of United Kingdom outlets with CC-News coverage falls from 13 in 2022 to only one in 2026, while Germany declines from 15 outlets to four and Australia from 13 outlets to two. Similar reductions are observed in the total number of matched HTML pages.

classifying captions. The construction of the caption dataset and the rule-based classification procedure are described in Appendix A.7.

The CC-News sample contains 956,911 downloaded image occurrences, of which 230,773 are associated with a nonempty caption. After excluding captions that could not be reliably translated into English because of their original language, we remain with 215,638 analyzable captioned image occurrences.

Explicit disclosure of AI-generated imagery is rare. We identify only 170 captions explicitly stating that an image was AI-generated or naming a generative AI system, corresponding to 0.09% of analyzable captions. Broadening the definition to include ambiguous references to AI increases the estimated prevalence to 0.43% of analyzable captions (487 captions).

Figure 2: Estimated prevalence of AI disclosures in image captions over time.



Notes: This figure reports the annual weighted share of analyzable image captions classified into four disclosure categories. Strict AI-image disclosures identify captions explicitly stating that an image was generated using artificial intelligence or naming a generative AI image model. AI editing disclosures identify captions reporting AI-assisted editing of an image. Any AI disclosure combines strict AI-image and AI editing disclosures, while broad AI references include ambiguous references to artificial intelligence. Estimates are restricted to captions originally written in English or successfully translated into English. Weighted percentages use the inverse article-sampling probability associated with each figure occurrence.

Figure 2 reports annual disclosure rates between 2022 and 2026. Although disclosure becomes more frequent over time, the increase is driven primarily by broad references to AI rather than explicit disclosure that an image was AI-generated. Strict AI-image disclosures remain close to zero throughout the sample period, increasing only modestly after 2024 and never exceeding 0.75%.

Taken together, these results indicate that while AI-generated images are becoming increasingly common in mainstream online news, explicit disclosure of their use remains exceptionally uncommon. The combination of rising adoption and near-zero disclosure implies that readers are unlikely to know when they are viewing AI-generated news imagery. Appendix Table A.50 compares the most frequent one-word and two-word phrases in captions that disclose AI use

with those in the remaining captions.

Prevalence of AI-generated images by country. Appendix Table A.4 reports the share of image occurrences classified as AI-generated separately for each country in both datasets.

Substantial cross-country variation is evident in both datasets. In CC-News, the highest classification share is observed in India (17.73%), followed by the United States (4.90%) and the United Kingdom (4.88%), while the lowest shares are observed in Japan (0.77%), Kenya (1.55%), and Italy (1.67%). The corresponding GDELT VGKG estimates are more compressed: India (4.34%), Brazil (4.14%), and Nigeria (4.02%) have the highest classification shares, while Kenya (0.57%), Japan (0.90%), and South Korea (1.96%) have the lowest.

These differences may reflect variation in editorial practices, outlet composition, and the types of images appearing in news articles, as well as differences in the performance of the ensemble across countries. Nevertheless, images classified as AI-generated are observed in all seventeen countries in the analysis, indicating that the use of AI imagery is geographically widespread rather than confined to a small number of news markets.

Prevalence of AI-generated images by topic. We next examine whether the use of AI imagery varies across types of news. We assign articles to categories from the IPTC Media Topics taxonomy based on the structure encoded in news outlets' URLs. Appendix A.8 describes the topic assignment procedure, and Appendix Table A.5 reports the share of images classified as AI-generated across the resulting article topics.

Substantial variation is evident in both datasets, particularly in CC-News. In CC-News, the highest classification share is observed in religion and belief (55.68%), followed by education (23.54%) and science and technology (19.27%). By contrast, classification shares are considerably lower in human interest (1.55%), disaster, accident and emergency incident (1.61%), and crime, law and justice (1.91%). No images are classified as AI-generated in conflict, war and peace or labour.

Differences across topics are less pronounced in the GDELT VGKG sample. The highest classification shares are observed in health (6.07%), science and technology (4.68%), and weather (4.48%), followed by lifestyle and leisure (3.74%). At the other end of the distribution, shares are comparatively low in disaster, accident and emergency incident (0.20%), human interest (1.66%), and society (1.84%). No images are classified as AI-generated in conflict, war and peace. Estimates for some categories rest on few positive image occurrences and should therefore be interpreted cautiously.

Overall, the adoption of AI imagery is not uniform across news content. Classification shares are relatively high in softer topics where images may serve an illustrative or representational role, including arts, culture, entertainment and media, lifestyle and leisure, and science and

technology. By contrast, shares tend to be lower in harder, event-driven topics such as crime, law and justice, and disaster, accident and emergency incident. Appendix Table A.6 reports the corresponding rates of AI disclosure in captions by topic.

Prevalence of AI-generated images by outlet type. We next examine whether the prevalence of AI-generated images varies systematically with the size of an outlet's online audience. We measure outlet readership using the Reuters Institute *Digital News Report* measure of weekly online use and match each observation to the corresponding outlet and survey year. Appendix Figure A.1 plots AI-image prevalence against weekly online readership.

The relationship is weak in both datasets, although the fitted trends differ slightly: in CC-News the share of AI images declines modestly with readership, while in GDELT VGKG it increases modestly. Neither relationship is pronounced.

Because both readership and the use of AI-generated imagery may evolve over time, Panel (b) residualizes both weekly online readership and the share of images classified as AI-generated with respect to year fixed effects. The fitted relationships remain largely unchanged, slightly negative in CC-News and slightly positive in GDELT VGKG, indicating that the weak association in the raw data is not driven by common changes in readership and AI-image classification shares over time.

AI-generated imagery is therefore not confined to a particular tier of the news market, and readers are unlikely to avoid it simply by consuming news from the largest and most established outlets.

3 Experimental Design

We run two controlled experiments in which respondents read the same news article(s), varying only the accompanying image and its source label. Both experiments share a common core structure: participants from the UK general population recruited on Prolific read an article from *The Conversation* paired with one of several image conditions, and indicate whether they wish to receive the outlet's newsletter. We begin by describing the shared features of the design before turning to how the two experiments differ.

3.1 Common Design Elements

Recruitment and sample quality. Participants were recruited from the UK general population on Prolific, with quotas on age, gender, and political affiliation to approximate the UK adult population. 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. Prolific's built-in speed check was enabled to exclude respondents completing the survey unreasonably fast. The median completion time in both experiments was eight minutes. Table C.2 compares observable sample characteristics with UK population benchmarks: both samples are broadly representative, though somewhat more educated and with higher average income.

Article stimulus. In both experiments participants read a real news article from *The Conversation*, a reputable UK academic-led outlet whose identity is made salient at the outset. Importantly, while reading the article participants are aware of the subsequent possibility to subscribe to the outlet's weekly newsletter, so that evaluations are clearly attributed to the outlet.

Treatment manipulation. The core manipulation varies (i) whether the accompanying picture is an original wire-service photograph or a closely matched AI-generated version, and (ii) whether a source label is displayed. In the control conditions, the original photograph carries a wire-service credit (e.g., *EPA-EFE/Caroline Brehman*) or is shown unlabelled. In the AI conditions, the AI-generated image is accompanied by a disclosure label (*"AI-generated image (created using OpenAI's DALL·E)"*) or no label. We list many examples of the labels across different platforms in Figure D.1. Full instructions are available in Appendix F.

We emphasize that the AI images are deliberately constructed to be content-equivalent to the wire-service originals: composition, subject placement, lighting, and semantic content are held constant, and in Experiment 2 a two-step protocol separates scene description from generation precisely to fix the depicted content while varying only photorealism. The images contain no misinformation and introduce no bias relative to the photographs they replace. This is intentional. Our object of interest is the consequence of AI origin, not of any content distortion that synthetic imagery can also carry. Allowing the AI images to differ in accuracy, framing, or evident quality would reintroduce exactly the confound that prevents existing studies from separating reactions to "AI" from reactions to "bad content." Holding content fixed means that any treatment effect on demand or perceived quality — and, critically, any penalty applied to authentic photographs that respondents misclassify as AI — is attributable to beliefs about provenance rather than to the image's informational properties. The design thus trades dramatic stimuli for clean identification of the channel that the rising, largely undisclosed use of routine AI imagery in mainstream outlets (Section 2) makes policy-relevant.

Demand for news (primary outcome). Our primary behavioural outcome is newsletter subscription (yes/no). 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*. We follow Chopra et al. (2022, 2024) and treat newsletter sign-up as a compact, incentive-compatible engagement measure. Email newsletters are a common

channel for news consumption (Newman et al., 2020), and the decision is behaviourally costly: subscribing means receiving weekly messages for a month, so it is not a free expressive gesture. Two pieces of evidence support the measure's external validity. First, Chopra et al. (2024) show in a dedicated validation experiment that respondents who sign up for a newsletter have a 25.9% higher incentivised willingness to pay (0.41 SD) for a twelve-month subscription to the underlying outlet ($p = 0.003$). Second, Alabrese et al. (2025), using the same elicitation, recontacted subscribers with the actual newsletter link and find that 86% of the 440 recontacted respondents clicked through, and that sign-up correlates significantly with stated willingness to read content from the source. Sign-up therefore reflects genuine intended consumption rather than costless assent.

Manipulation check and recognition coding. After the primary outcome, we ask: "*Based on the sample article you just saw, was the image AI-generated?*" (Yes/No). We construct a recognition indicator:

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

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). We report results using a standardized composite quality index (average of the three z-scored items).

Debrief. To avoid deception, respondents are truthfully told that the text is from a real article. Because possible AI image use is not mentioned ex-ante, participants are debriefed at the end of the survey with full information about the synthetic picture and the objective of the study. The content was benign and posed minimal risk.

3.2 Experiment 1: Single Article

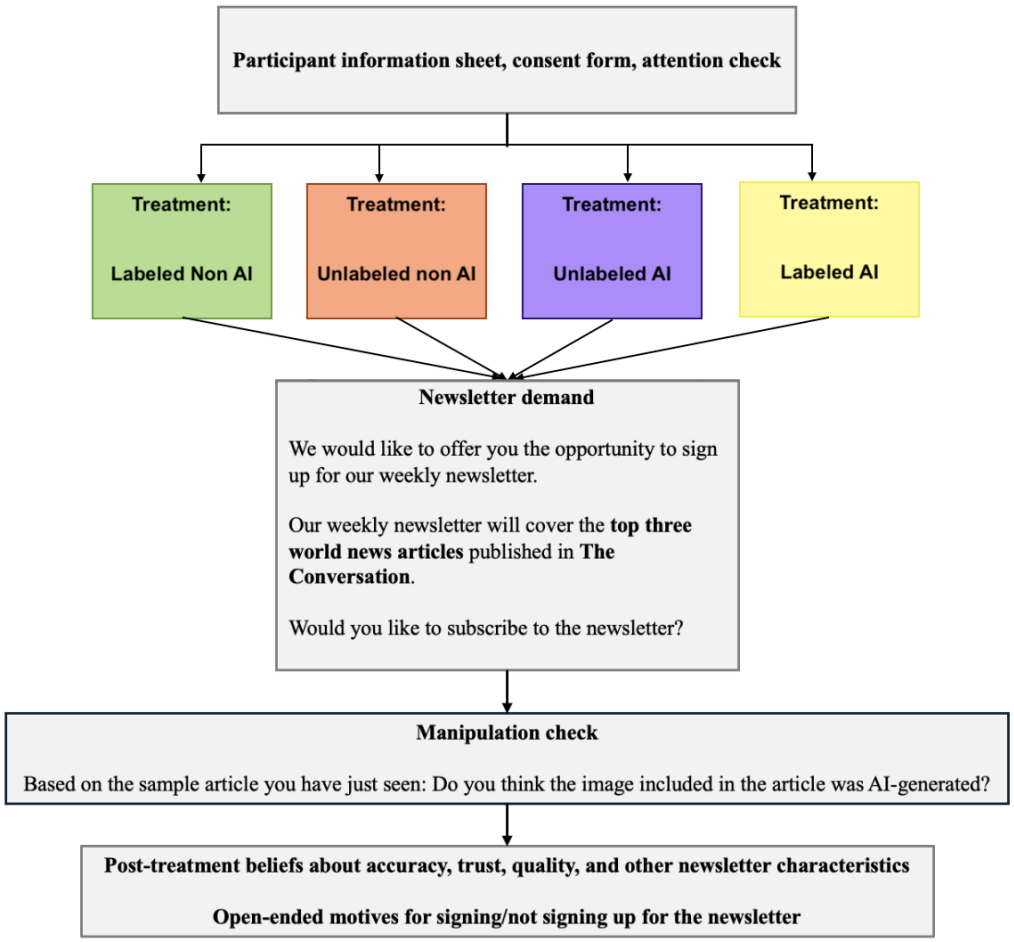
Experiment 1 was pre-registered on AsPredicted (n. 251218).[21] Responses were collected between October 14 and 19, 2025. After pre-registered exclusions, the final sample consists of 2,870 respondents (target $N \approx 3{,}000$). Respondents were paid £1.5 (over £11 per hour).

Participants were randomly assigned to one of **four image conditions**: Unlabelled Control, Labelled Control, Unlabelled AI, and Labelled AI. Figure 3 provides an overview. The experimental stimulus is a single real news article from *The Conversation* discussing recent protests

[21] Pre-analysis plan available at: aspredicted.org/xw49-bcwx.pdf.

in California.[22] For a UK sample, the topic is politically relevant yet psychologically distant, encouraging engagement while limiting strong personal identification or sacred-value reactions. The AI image was generated using DALL·E; composition, lighting, and subject placement were designed to closely mirror the original (Figure 4b). Full generation prompts are in Appendix E. Table C.3 confirms balance across conditions.

Figure 3: Experiment Overview



[22] Original article available at this link.

Figure 4: Treatment Manipulation (Experiment 1)



(a) Original (Control with/without credit) **(b)** AI Image (Treatment with/without label)

3.3 Experiment 2: Multiple Articles and Photorealism Variation

Experiment 2 was pre-registered on OSF (e5ncv). After pre-registered exclusions, the final sample consists of 4,196 respondents. Respondents were paid £2 (over £14 per hour). Table C.12 confirms balance across conditions.

Experiment 2 extends Experiment 1 in two directions. First, each respondent views exactly one article sampled with equal probability from a fixed set of **six news articles** covering diverse topics (Business & Economy, Politics & Society, Arts & Culture, Education, Environment), all from *The Conversation*. These topics broadly correspond to the categories in which AI-generated imagery is actually used in mainstream reporting. In the audit of Section 2, education and arts, culture, entertainment and media are among the categories with the highest measured AI-image prevalence, while economy, business and finance and arts and culture are among the largest topics by volume in both archives. Politics is deliberately retained even though it records among the lowest prevalence, so that the stimulus set spans the range of observed adoption rather than only its upper tail. Because our research question concerns responses to image authenticity rather than heterogeneity across article content, we treat the article dimension as a nuisance parameter and pool observations across articles with article fixed effects.

Second, it introduces **randomised variation in AI photorealism**. For each article, AI images were generated using a two-step protocol that separates scene description from image generation (see Appendix E). This yields two AI variants: a highly photorealistic version (*Hard AI*, closely resembling a professional news photograph) and an intentionally less photorealistic version (*Easy AI*, with visual cues making AI generation more apparent). Semantic content and composition are held constant across variants. Because this variation is exogenously assigned, it shifts the probability that respondents identify the image as synthetic without altering image origin or label, providing a clean instrument for the recognition channel.

Participants were randomly assigned to one of **six image conditions**:

1. **Control, no label:** real photograph with no credit or disclosure.
2. **Control, label:** real photograph with a conventional wire-service credit.
3. **Hard AI, no label:** highly photorealistic AI image with no disclosure.
4. **Hard AI, label:** same photorealistic AI image with an AI disclosure.
5. **Easy AI, no label:** non-photorealistic AI image with no disclosure.
6. **Easy AI, label:** same non-photorealistic AI image with an AI disclosure.

Conditions 1–4 replicate the core origin $\times$ label design of Experiment 1 on a richer image set; conditions 5–6 provide the exogenous variation in recognizability. The newsletter subscription decision is elicited *before* any recognition question, avoiding priming of image scrutiny.

4 Results

Across both experiments, AI recognition is poor and both outlet perceptions and demand behaviour hinge on whether people *believe* the image is AI-generated. We first document the core findings from Experiment 1 and show that they replicate and extend in Experiment 2. We then use Experiment 2's photorealism instrument for causal identification of the recognition channel.

4.1 Manipulation: Poor AI Recognition and Weak Labelling Effects

Figure 5a reports AI recognition rates (Experiment 2; Experiment 1 in Appendix Figure D.2). In Experiment 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 C.4) confirms that the AI $\times$ 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.

In Experiment 2, the pattern replicates and is amplified by the photorealism manipulation. Among respondents assigned the less photorealistic AI image, correct origin recognition rises sharply to 81.2% in the unlabeled easy-AI arm and 86.5% in the labeled easy-AI arm (both $p < 0.001$ relative to the unlabeled control). By contrast, only 44.1% of respondents in the

unlabeled hard-AI arm correctly identified it as AI, closely mirroring the 46% and 52% observed in Experiment 1's unlabeled and labeled AI arms. These results confirm that photorealism is the primary driver of recognition, while labels offer only modest correction regardless of image type a gap of roughly 37 percentage points between easy-AI and hard-AI conditions (Table C.13).

Overall, 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, indicating that respondents spent time with the stimulus but still struggled to recall 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).

4.2 Engagement: No Average Effect of AI or Labels

Experiment 1. Figure D.3a (Appendix) and Table C.5 show that newsletter demand is essentially flat across conditions. 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 produces a small and insignificant increase (39.3%), and combining AI with a label yields 37.2%. Neither AI, nor Label indicators, nor their interaction is statistically different from zero, and the joint test is also clearly insignificant.[23]

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.

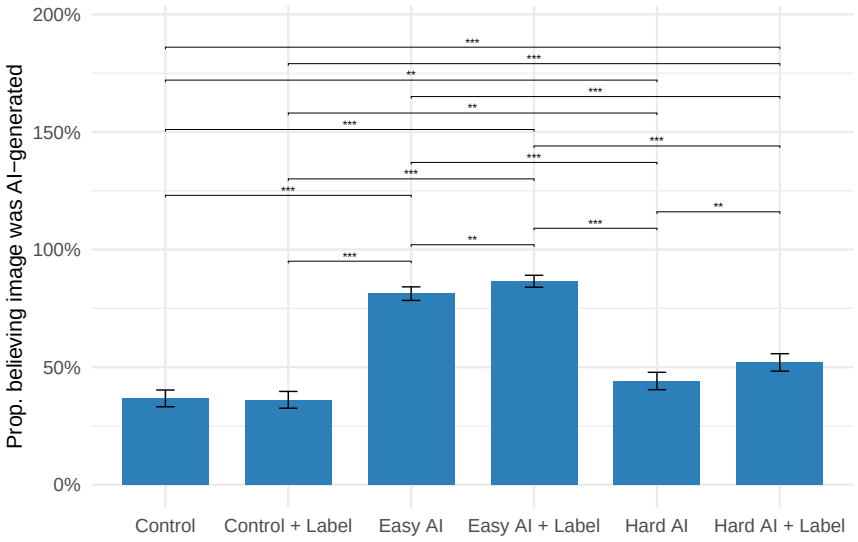
Experiment 2. Figure 6a and Table C.15 replicate the aggregate null across six articles. Baseline newsletter demand in the unlabeled control is 35.7%, and estimated differences across all other arms are small and statistically indistinguishable from zero: -1.5 p.p. in the labeled control, $+1.7$ p.p. in the unlabeled easy-AI arm, -2.4 p.p. in the labeled easy-AI arm, -0.6 p.p. in the unlabeled hard-AI arm, and -2.0 p.p. in the labeled hard-AI arm (none significant; joint test also clearly insignificant). These results hold with and without article fixed effects, confirming the null is not driven by compositional differences across articles.

Substantively, holding the article text and outlet constant, swapping a real photo for an AI image does not shift short-run demand.

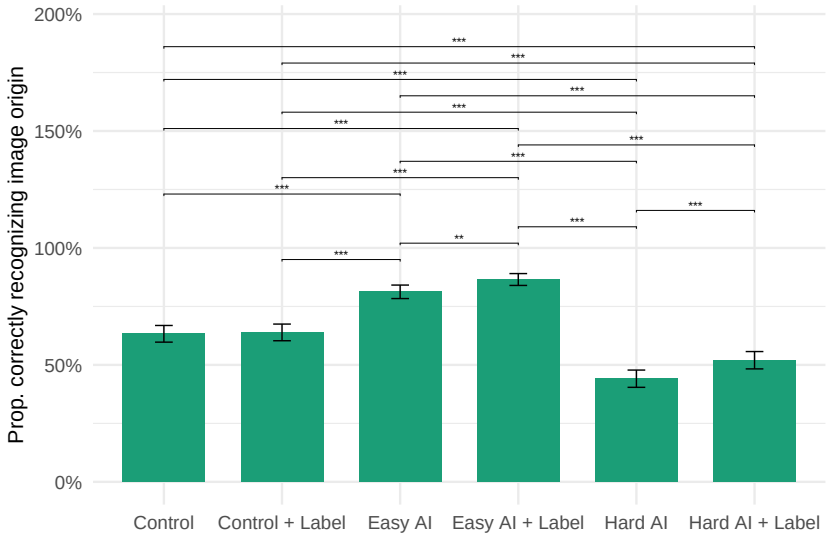
[23] Crucially, while inspecting the article, respondents were aware that they would later have the option to subscribe to the outlet's weekly newsletter. They also spent over 90 seconds on the article page, with no systematic differences across arms (Table C.3), indicating balanced and sufficient attention. As a robustness check (not shown), we re-estimated all outcome effects controlling for reading time; results are unchanged.

Figure 5: 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 (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$). Only significant pairwise comparisons are reported. Experiment 1 results appear in Figure D.2.

4.3 Heterogeneity: High Demand with Unnoticed AI, Low if Recognised

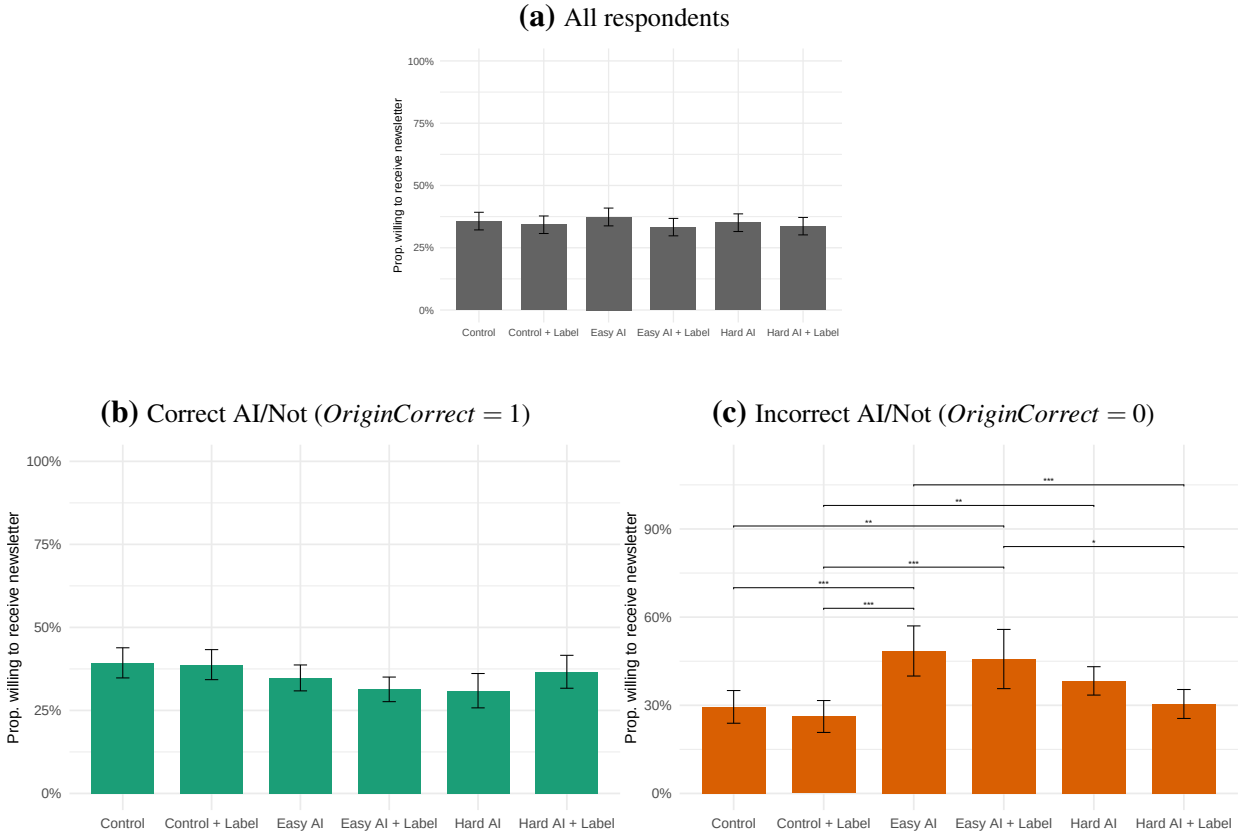
Experiment 1. Among respondents who *did not* correctly identify the image source, newsletter demand in an AI image condition is approximately 10 percentage points higher (0.081–0.105,

$p < 0.01$) than with real images. Among those who *did* correctly identify the image source, newsletter demand is instead lower by a similar amount in the case of the AI image (-0.097 to -0.090, $p < 0.01$) as compared to the real image (Table C.6; Figure D.3 in Appendix). The non-overlapping confidence intervals visually reinforce the sign reversal.

Labeling does not close the gap: the three-way AI $\times$ Label $\times$ OriginCorrect interaction is small and statistically insignificant (Table C.5). What matters is not whether the image is actually AI, but whether respondents believe it is. Figures D.5a and D.5b confirm this: stratifying demand solely by whether respondents believed the image was AI-generated (regardless of accuracy), treatment assignment does not move demand conditional on belief, while the gap between belief groups is stark.

Experiment 2. The same sharp heterogeneity replicates across six articles and twelve images, now with exogenous identification from the photorealism manipulation. Among respondents who did *not* correctly recognize the image origin, newsletter demand rises substantially: $+19.5$ p.p. in the unlabeled easy-AI arm ($p < 0.001$), $+16.4$ p.p. in the labeled easy-AI arm ($p < 0.001$), and $+8.9$ p.p. in the unlabeled hard-AI arm ($p < 0.05$). Among respondents who *did* correctly identify the image as AI-generated, the pattern reverses sharply: the interactions are negative and significant for the unlabeled easy-AI arm (-0.240, $p < 0.001$), the labeled easy-AI arm (-0.244, $p < 0.001$), and the unlabeled hard-AI arm (-0.174, $p < 0.001$) (Table C.15; Figures 6b and 6c).

Figure 6: Newsletter Demand by Treatment and AI Recognition Accuracy



Notes: Each panel shows the share of respondents demanding the newsletter by treatment arm (Experiment 2, $N = 4{,}196$). 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 (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$). Only significant pairwise comparisons are reported. Experiment 1 replication appears in Figure D.3.

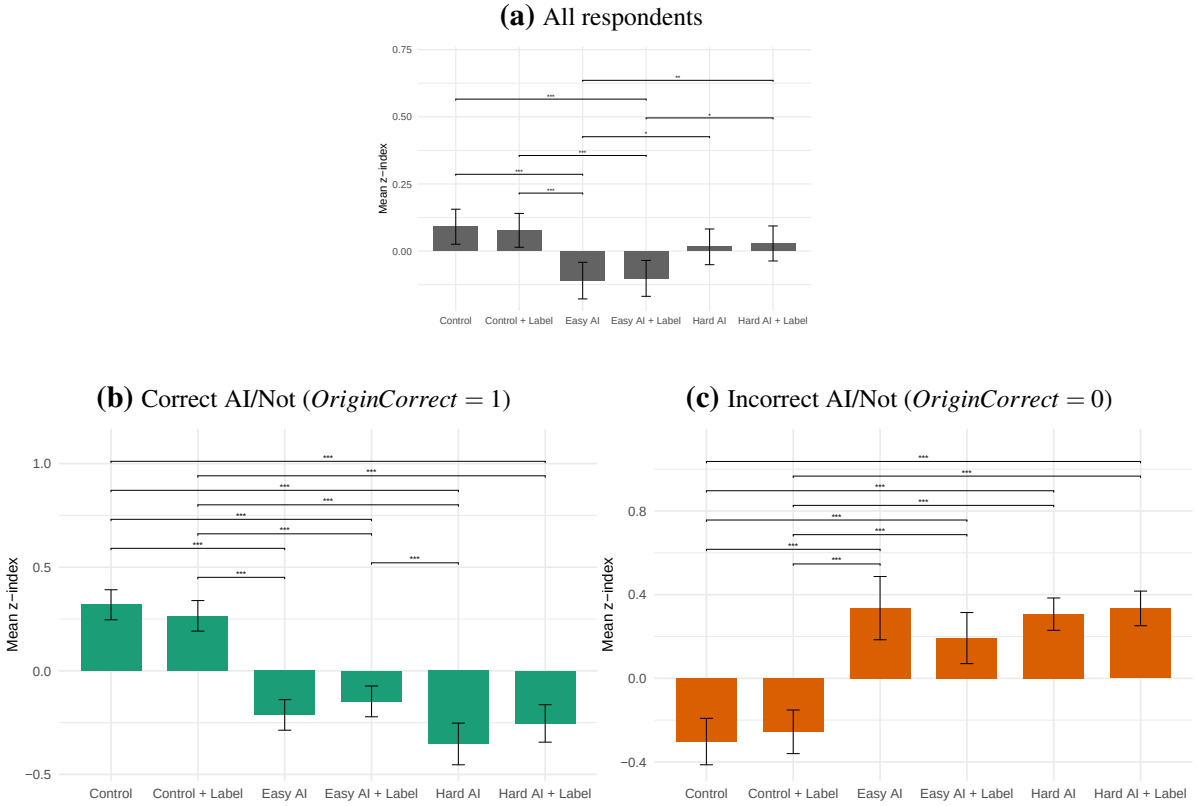
Two implications follow. *First*, the image source (AI vs. real) does not move average engagement unless users internalise it; recognition is crucial. *Second*, disclosure helps at the margin but is not a remedy, consistent with mixed evidence on labeling regimes (Altay and Gilardi, 2024; Wittenberg 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.

4.4 Mechanism: Credibility and Engagement Move Together

Experiment 1. Perceived news source quality mirrors the heterogeneous demand findings (Figure D.4 in Appendix; Table C.7). Pooling all respondents, average credibility barely differs across treatment arms. Among respondents who *did not* correctly identify the image source, the AI image raises perceived credibility by roughly 0.6–0.7 standard deviations relative to the control arm ($p < 0.001$). Among those who *did* recognise the image as AI, the pattern reverses: the $AI \times OriginCorrect$ interactions are strongly negative (-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. As with demand, Figures D.5c and D.5d show that quality perceptions are flat within belief groups and differ sharply between 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.

Experiment 2. The mechanism replicates with comparable magnitude across six articles. Among respondents who did *not* correctly identify the image origin, all AI conditions exhibit higher perceived quality by roughly 0.5–0.6 standard deviations (Figure 7b). Among respondents who *did* correctly identify the image as AI-generated, the $AI \times OriginCorrect$ interactions in Table C.16 are all large and negative (-0.96 to -1.29 SD, all $p < 0.001$), implying net credibility losses of approximately -0.3 to -0.7 SD. This pattern holds uniformly across both easy-AI and hard-AI conditions: once AI images are recognized, the credibility penalty is comparably large regardless of whether the image was easy or hard to detect as synthetic. Labels offer no meaningful attenuation.

Figure 7: Perceived Outlet Quality by Treatment and AI Recognition Accuracy



Notes: Each panel shows the average perceived outlet quality by treatment arm (Experiment 2, $N = 4{,}196$). 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 (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$). Only significant pairwise comparisons are reported. Experiment 1 replication appears in Figure D.4.

The joint movement of credibility and sign-ups is consistent with models in which perceived outlet quality drives news demand (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.

s a complementary test in Experiment 1, we treat reported AI recognition as endogenous and instrument it using the randomized AI $\times$ Label interaction (Table C.8). The first stage confirms the interaction raises the likelihood of reporting the image as AI-generated by about 0.10 ($p < 0.01$). In the second stage, the 2SLS estimate for perceived outlet quality is negative and statistically significant ($p < 0.05$), while the estimate for newsletter demand is negative but imprecise, consistent with the small reduced-form effects of AI $\times$ Label. The excluded instrument is, however, only modestly strong (robust first-stage F of 7.2–8.4, below the conventional threshold of 10); we therefore report weak-instrument robust Anderson–Rubin confidence sets alongside the 2SLS estimates in Table C.8 and treat the Experiment 1 IV as corroborative evidence only.[24] The paper's causal claims rest on the Experiment 2 photorealism instrument (first-stage $F > 90$; Section 4.5), to which we now turn.

We also test several alternative explanations (Table C.9). AI images do not increase perceived political bias or confusion about the text. Average trust in the researcher remains above "somewhat trustworthy" in every arm, and systematic declines appear only when respondents recognise the image as AI mirroring the credibility penalty applied to the news outlet rather than signalling broad experiment distrust. Additional checks (guessed study purpose in Table C.10; open-ended newsletter reasons in Table C.11) confirm that respondents did not infer our key hypothesis and were responding to content rather than second-guessing the experiment.

4.5 Causal Identification: The Role of Recognition

The heterogeneity results above establish that demand and quality move with beliefs about image authenticity. However, the recognition indicator (*OriginCorrect*) is elicited after treatment exposure and may reflect post-treatment selection rather than a causal belief mechanism. Experiment 2 addresses this directly with the photorealism instrument.

First stage. The photorealism manipulation is a strong instrument. The robust partial F-statistic for the excluded instrument reaches 171.1 when the endogenous variable is *Manipulation* and 91.2 when it is *OriginCorrect* (Table C.17), both far exceeding conventional weak-instrument thresholds. Respondents assigned the less photorealistic AI image are substantially more likely to identify it as synthetic.

[24] The Anderson–Rubin sets are wider than the 2SLS intervals but exclude zero for the quality index and include it for demand, matching the 2SLS pattern. [Add AR rows to Table C.8.]

Newsletter demand. The 2SLS estimates for newsletter demand are small and statistically indistinguishable from zero: the instrumented coefficient on *Manipulation* is $+0.007$ (s.e. 0.036) and on *OriginCorrect* is $+0.015$ (s.e. 0.050). These estimates are consistent with the aggregate null arising from symmetric, offsetting belief updates rather than the absence of any belief-driven mechanism.

Perceived outlet quality. The causal evidence is sharper for perceived outlet quality. The instrumented coefficient on *Manipulation* is -0.366 (s.e. 0.066, $p < 0.001$) and on *OriginCorrect* is -0.348 (s.e. 0.091, $p < 0.001$), both robust to the inclusion of controls (Table C.17). These provide clean causal evidence that recognition not mere exposure to an AI image is the operative mechanism depressing credibility: correctly identifying the image as AI-generated reduces perceived outlet quality by approximately 0.35–0.37 standard deviations.

Exclusion restriction. The photorealism instrument identifies the recognition channel only if reduced realism affects demand and perceived quality exclusively through the recognition judgment, and not directly, for example, through aesthetic displeasure or an inference of low production effort. Three patterns in the data support exclusion. First, conditional on the recognition judgment, quality perceptions are statistically similar across easy-AI and hard-AI arms (Table C.16; Figure D.6): if lower photorealism depressed quality directly, easy-AI respondents should rate the outlet worse than hard-AI respondents *within* the same belief group, which they do not. Second, and most tellingly, respondents who failed to recognise the easy-AI image as synthetic rate outlet quality as high as indeed slightly higher than the control group, despite having viewed the least realistic stimulus; a direct aesthetic channel would push these ratings down, not up. Third, perceived article complexity and political bias are unaffected by the instrument (Table C.18), ruling out a broad stimulus-driven scepticism. A residual caveat remains: we cannot separate the recognition judgment itself from evaluative reactions that occur only *conditional on* recognising the image as AI. The IV estimand should therefore be read as the total effect of coming to believe the image is AI-generated, inclusive of any aesthetic re-evaluation that this belief triggers, which is the policy-relevant object for an outlet deciding whether recognisable AI use is costly.

Table C.18 further instruments *OriginCorrect* with the photorealism manipulation and breaks the quality index into its constituent dimensions. The effect is visible across all three quality dimensions: perceived accuracy falls, trustworthiness declines, and overall quality deteriorates. By contrast, perceived political bias and article complexity are unaffected, ruling out the possibility that recognition triggers a more general scepticism about article content. The one positive effect is on entertainment value: recognition reduces the probability that respondents rate the article as not entertaining, consistent with correctly identified AI imagery attracting curiosity value without translating into higher engagement.

These IV results complement those from Experiment 1, where the instrument was the randomized AI $\times$ Label interaction. Both instruments deliver the same conclusion: recognition causally depresses outlet credibility, while the null effect on demand arises from symmetric, offsetting belief updates.

5 Structural Exercises

The reduced-form evidence in Section 4 establishes three facts: (i) average newsletter demand is unaffected by AI imagery across all treatment arms; (ii) this aggregate null conceals sharp, belief-driven heterogeneity—a belief-conditional demand gap of 6.2 percentage points in the pooled Experiment 2 conditional means (and roughly 10 points in Experiment 1's regression-based heterogeneity); and (iii) the causal quality penalty from recognition is 0.35 standard deviations. These estimates are internally valid but leave open three policy-relevant questions that the reduced-form design cannot answer. How long does the aggregate null persist as audiences become better at detecting AI imagery? At what audience scale does AI adoption become unprofitable for an outlet? And what does the current noisy-disclosure regime cost, and what label policy would a social planner prefer?

We address these questions through three complementary structural exercises. For internal consistency, *all* exercises use a single set of moments from Experiment 2 (Table B.4): the belief-conditional demand and quality means $D^{\pm}$ and $Q^{\pm}$, and the arm-level attribution rates. Experiment 2 is the appropriate anchor because it is the larger, multi-article design, it contains the exogenous photorealism variation that identifies the recognition channel, and it is the benchmark to which the journalist survey in Section 6 is incentivised.[25] The first two exercises are back-of-the-envelope calculations that use the mixing equation (9) to translate the experimental estimates into quantitative predictions. The third recovers the structural parameters by a calibrated closed-form minimum-distance procedure, subjects the model to two auxiliary specification tests, and uses the estimated demand system to compute counterfactual consumer surplus under alternative AI-adoption and disclosure regimes. Full derivations, tables, and bootstrap confidence intervals appear in Appendix B.11–B.15.

5.1 When Does the Aggregate Null Dissolve?

The aggregate null arises from a mechanical near-cancellation: in every treatment arm a substantial share of respondents believe the image is AI-generated and a substantial share do not, and the

[25] The pooled Experiment 2 belief-conditional demand gap ($D^{+} - D^{-} = 6.2$ p.p.) is smaller than the arm-specific interaction coefficients in Table C.15 and than the Experiment 1 regression-based gap of roughly 10 p.p. We deliberately use the conservative pooled conditional means throughout, so that the back-of-the-envelope exercises and the structural estimation share one demand system; using the Experiment 1 gap would roughly scale the demand-side losses below by a factor of 1.6.

two groups' demand responses largely offset. As AI-detection tools improve and media-literacy interventions diffuse, this balance will erode. We quantify the critical detection threshold using the mixing equation (9) and the directly estimated belief-conditional demand rates from Experiment 2: $D^- = 32.2\%$ (demand when the reader believes the image is AI) and $D^+ = 38.4\%$ (demand when the reader believes it is real), a gap of 6.2 percentage points (Table B.4).

Parameterising joint improvements in AI-detection and real-image authentication by $\lambda \in [0,1]$, with recognition rates anchored at the Experiment 2 unlabeled-arm attribution rates ($\alpha_{\text{AI}}(0) = p_L = 0.441$ for photorealistic AI images; $\alpha_{\text{REAL}}(0) = p_H = 0.367$ for authentic photographs), the average demand gap between the AI-image arm and the real-image arm evolves as

$$\text{Gap}(\lambda) \;=\; -0.005 - 0.057\,\lambda, \tag{1}$$

which is mildly negative already today—because photorealistic AI is attributed to AI slightly more often than authentic photographs—and reaches its ceiling of -6.2 percentage points under full information. As a validation, the model-implied current gap of -0.5 p.p. closely matches the observed -0.6 p.p. difference between the unlabeled hard-AI and unlabeled control arms (Table C.15).

When does this gap become statistically detectable? At Experiment 2's total sample size ($N = 4{,}196$) split evenly between AI and real arms, the minimum detectable effect is 4.1 percentage points (80% power, $\alpha = 0.05$, two-sided).[26] Setting $|\text{Gap}(\hat{\lambda})| = 0.041$ yields

$$\hat{\lambda} \approx 0.64. \tag{2}$$

At $\hat{\lambda}$, the fraction of readers correctly identifying photorealistic AI images would need to rise from 44% today to approximately 80%, while the false-positive rate for authentic photographs would need to fall from 37% to about 13% (Table B.2).

Two features of this calculation are policy-relevant. First, reducing false positives for authentic photographs matters as much as improving AI detection: the currently high false-positive rate keeps control-group demand artificially depressed and compresses the observable gap. Authentication standards such as the C2PA Content Credentials protocol therefore contribute as much to the eventual emergence of a detectable demand penalty as AI-detection classifiers do. Second, even at full information ($\lambda = 1$), the aggregate demand penalty is bounded by the belief-conditional gap of 6.2 percentage points, a finite and, on many metrics, modest ceiling on the demand consequences of AI visual adoption. The causal quality penalty of 0.35 standard deviations (Table C.17) operates through a separate reputational channel and is likely to dominate in the long run.

[26] MDE $= 2.8 \times$ SE with SE $= \sqrt{2\bar{p}(1-\bar{p})/n}$, $\bar{p} = 0.35$, $n = 2{,}098$ per group. At Experiment 2's actual per-arm sizes (≈ 700), pairwise arm comparisons have an MDE of 7.1 p.p.—larger than even the full-information gap of 6.2 p.p. One reason the arm-level null is so robust in our data is thus that the belief-conditional demand gap is genuinely modest, not merely undetected.

5.2 Outlet Breakeven Cost of AI Visual Adoption

From the outlet's perspective, the relevant calculation is whether the per-image cost saving from AI adoption exceeds the recognition-driven engagement loss. Using equation (1), AI adoption breaks even when the cost saving δ satisfies

$$\delta > \big(0.005 + 0.057\,\lambda\big) \cdot m \cdot R_{\text{eff}}, \tag{3}$$

where m is the lifetime revenue per newsletter subscriber and R_{eff} is the effective audience per article. Rearranging, the breakeven detection level is

$$\lambda^*(\delta, m, R_{\text{eff}}) = \frac{\delta}{0.057 \cdot m \cdot R_{\text{eff}}} - 0.08. \tag{4}$$

We anchor the parameters to industry benchmarks as before: editorial image licensing of $\delta = £150$ and subscriber lifetime value of $m = £15$ (full sourcing in Appendix B.12). Substituting yields $\lambda^* \approx 175/R_{\text{eff}} - 0.08$.

Two implications follow, both sharper than under a zero current gap. First, because $\text{Gap}(0) < 0$—photorealistic AI already carries a small recognition-driven demand cost at today's detection rates—there is a critical audience scale at which AI imagery is unprofitable *now*: setting $\lambda^* = 0$ gives $R^* = \delta/(0.005 \cdot m) \approx 2{,}200$ effective readers per article. For a major publisher with 10,000 readers per article, the current engagement loss is roughly £680 per article against a £150 saving; for a micro-outlet with 100 readers, the loss even under full information (£92) never reaches the saving, so AI imagery remains profitable at any detection level. Second, the asymmetry is structural: the cost saving δ is a fixed per-image amount, while the engagement loss scales linearly with audience size. AI imagery is therefore a rational editorial choice for small and specialist outlets and an increasingly poor one for any outlet with meaningful traffic, a conclusion that follows directly from the demand estimates, independently of any ethical or editorial judgment about AI use.

An observable implication follows. Large outlets should retreat from AI visuals substantially earlier than small outlets in response to rising AI-detection capability, not because of greater ethical sensitivity, but because the cost-benefit calculation turns negative sooner. This prediction is testable against newsroom adoption data as detection tools diffuse.

5.3 Structural Parameter Recovery and Welfare Analysis

This subsection takes the model of Appendix B to the data: we recover its parameters, test its restrictions, and use the estimated demand system to evaluate the welfare effects of AI adoption, detection, and disclosure. Because some parameters are implausible if read literally, the welfare analysis rests only on directly estimated objects.

Parameters and how they are recovered. The model has four structural parameters: the prior π (the share of readers who believe the outlet is of high type before seeing the image); the normalized value spread $\Delta v = (v_H - v_L)/c_{\max}$ between high- and low-type newsletters; the normalized floor demand $v_L/c_{\max}$; and the quality sensitivity b, the slope of the quality z-index on the posterior belief μ. Two further inputs are taken directly from the data rather than estimated: the false-positive recognition rate $p_H = 0.367$ (the share of readers attributing an authentic photograph to AI, from the control arm) and the true-positive rate $p_L = 0.441$ (the share attributing a hard-to-detect AI image to AI).

With p_H and p_L fixed, recovery is exact. The four moment conditions g_1–g_4—mean demand and mean quality among readers who believe the image is AI (D^-, Q^-) and among those who believe it is real (D^+, Q^+)—pin down $(\Delta v, v_L, b)$ in closed form for any given π. The fourth condition adds no information: it holds automatically once b is set, because the quality index is standardized on the same sample used to construct the average posterior $\bar{\mu}$. The remaining degree of freedom, the level of π, is fixed by normalizing the floor demand to zero ($v_L = 0$), which does not affect any of the gaps the model identifies. The system is thus just-identified, with no overidentification test to run; the model does, however, imply two further restrictions, which we test below.

The recovered values are $\hat{\pi} = 0.439$, $\widehat{\Delta v} = 0.816$, and $\hat{b} = 7.10$, with bootstrap standard errors of 0.105, 0.165, and 0.573 (Appendix B.13). The value spread is large: the gap between a high- and a low-type newsletter amounts to more than 80% of the maximum subscription cost, so readers care substantially about the quality of the outlet.

Interpreting $\hat{b}$. The quality sensitivity should not be read literally. Because p_H and p_L are close (0.367 versus 0.441), recognising an image as AI is weak evidence about the outlet, so the model's posteriors barely move ($\mu^+ - \mu^- = 0.075$). Explaining the observed 0.54 SD quality gap with so small a belief change forces an extreme slope. Taken at face value, $\hat{b} = 7.1$ implies that a reader who became certain the outlet was low type would lower perceived quality by seven standard deviations—which is implausible. A more natural reading is that respondents do not perform the small Bayesian update the model prescribes. Instead, they react categorically to the belief "this image is AI," applying a roughly fixed penalty whenever they hold it. The data offer only weak support for the model's diagnosticity assumption (Assumption B.2), which requires AI images to be recognised more often than authentic ones: it holds in Experiment 2's unlabeled arms (44.1% versus 36.7%) but fails marginally in Experiment 1's (45.8% versus 47.2%), where the photorealistic AI image was attributed to AI slightly *less* often than the authentic photograph. We therefore treat the Bayesian structure as a convenient representation of belief-conditional behaviour, not a literal account of updating, and build the welfare analysis on $D^{\pm}$, $Q^{\pm}$, and arm-level attribution rates rather than on π, $\mu^{\pm}$, or b.

Specification tests. The model's central testable restriction is within-group homogeneity: once we condition on whether a reader believes the image is AI (recognition judgment r), demand and quality should not depend on which treatment arm produced that belief. Arm-by-arm F-tests (Appendix Table B.6) do not reject this for newsletter demand among readers who believe the image is AI ($F(5, 2349) = 2.10$, $p = 0.062$), but do reject it among those who believe it is real ($F(5, 1835) = 3.40$, $p = 0.005$). The source of the rejection is easy to identify. Readers shown an obviously synthetic image who nonetheless concluded it was real (the easy-AI arm with $r = 0$) report newsletter demand of 47.4%, well above the 39.1% of believe-real readers in the control arm. This hidden-AI bonus is too large to be consistent with homogeneity. The second restriction, cross-equation proportionality, $\Delta v/b = (D^+ - D^-)/(Q^+ - Q^-)$, holds on the pooled sample by construction but varies considerably across arms, from -0.105 in the labeled hard-AI arm to 0.424 in the labeled easy-AI arm (Appendix Table B.7). The model thus captures average behaviour well, while the label-by-photorealism interaction produces heterogeneity it does not fully capture.

Measuring welfare. Each disclosure or detection scenario implies a recognition rate α: the share of readers who come to believe the image is AI. Readers who hold that belief demand D^-; the rest demand D^+. Under a uniform subscription-cost distribution, per-reader consumer surplus is $\mathrm{CS}(\alpha) = \alpha(D^-)^2/2 + (1-\alpha)(D^+)^2/2$, so surplus falls as more readers recognise the image as AI. Two features matter. First, because the experiment involves no price variation, the upper bound $c_{\max}$ of the cost distribution is not identified. Surplus is therefore measured in units of $c_{\max}$ per reader, and we report changes as percentages of baseline surplus, which do not depend on $c_{\max}$. Outlet profit is computed separately, in pounds, from an externally calibrated subscriber value m; we never add the two together at an arbitrary exchange rate (Appendix B.15.1). Second, the measure uses only $D^{\pm}$ and the arm-level attribution rates, so it does not depend on π or $\hat{b}$.

Welfare effects of AI adoption and detection. All comparisons are relative to a baseline in which the outlet publishes authentic photographs. Three findings stand out. First, switching entirely to hard-to-detect photorealistic AI images, without labels, reduces consumer surplus by 2.5% (95% bootstrap confidence interval $[-4.8\%, -0.5\%]$). This is the welfare cost of AI adoption at current detection rates, and it is statistically distinguishable from zero. Second, the cost rises steeply with how recognisable the imagery is. Easy-to-detect AI images reduce surplus by 14.8%, nearly six times as much, because far more readers recognise them and move into the low-demand state D^-. Third, we vary readers' detection capability λ, from $\lambda = 0$ (current detection) to $\lambda = 1$ (every reader recognises AI images). Surplus falls monotonically in λ, from 2.5% below baseline today to 21% below it under full detection. At the other extreme, if no reader ever recognised AI images, measured surplus would be 12.2% *above* baseline. Each ten-percentage-point improvement in detection lowers surplus by $0.0012\,c_{\max}$ per reader per article. Per 1,000 readers, this is £$1.21 \times c_{\max}$ per article, or roughly £6 to £61 for values of $c_{\max}$

between £5 and £50. The outlet additionally loses about £52 in revenue per article per 1,000 readers at $m = £15$.

What the welfare numbers measure. These are *decision-utility* measures: they value a subscription at what readers believe it is worth. Two caveats follow, developed formally in Appendix B.15.6. First, the 12.2% gain when no reader recognises AI images comes from a mistaken belief—readers value AI-illustrated content as if it were photographed—and would not count as a gain under experienced utility. Second, the diagnosticity assumption that makes a reader's downgrade rational presumes that, across outlets, AI use signals lower quality—a link our design deliberately removes by holding content and quality fixed. Readers' inferences may thus be sound about the market yet wrong about the outlet in front of them; in a rational-expectations equilibrium with no such link, the belief-conditional gap could not persist. The implausibly large $\hat{b}$ points instead to a direct, categorical distaste for recognised AI provenance. Under that reading, decision utility is the appropriate metric and our estimates stand as reported. The policy conclusions below hold under either reading.

Optimal disclosure policy. We next ask how intensively a planner should require AI labels, with label intensity $\gamma \in [0,1]$ and $\gamma = 1$ corresponding to mandatory disclosure. Labeling a hard-to-detect AI image raises the recognition rate by 7.9 percentage points, from 44.1% to 52.0%. At the estimated $D^{\pm}$, both consumer surplus and outlet profit fall as recognition rises, so the optimum is $\gamma^* = 0$: no mandatory labeling. Because both welfare components move in the same direction, this conclusion holds for any monetary value of $c_{\max}$ and any planner weighting of readers against the outlet (Proposition B.10); it does not depend on how the two components are scaled. The result is specific to settings like our experiment, where AI imagery leaves content quality unchanged: a label then pushes readers toward the pessimistic belief without conveying anything about the newsletter's value, reducing surplus without correcting any real information gap. If instead AI use *were* associated with lower editorial standards, as in Goldberg and Lam (2025), disclosure would correct a genuine quality difference, and a planner would find some labeling optimal.

An unambiguous gain: authenticating real images. One component of the current regime carries a welfare cost under any metric: the penalty that false positives impose on authentic photographs. For an outlet publishing only real images, eliminating the 36.7% false-positive attribution rate, for example through authentication standards such as C2PA, would raise consumer surplus by 12.2% of baseline (Appendix B.15.5). This is the same number as the gain when no reader recognises AI images, and for the same reason: in both cases every reader ends up believing the image is real. The difference lies in whether that belief is correct. For AI images it is mistaken, so the gain is illusory; for authentic images it is accurate, so correcting

false positives moves beliefs toward the truth *and* raises surplus. Authenticating real images is therefore the disclosure investment with clear welfare value, whereas the value of labeling AI images depends entirely on whether AI use is associated with lower quality.

6 Journalists' Beliefs

To assess external validity, we conducted a survey with 132 international journalists recruited via Prolific. The sample had an average age of 35.1 years and skewed female (59.8%). Respondents were spread across multiple countries, with the United States and the United Kingdom being the two most represented markets (35.1% and 21.1% respectively). In terms of outlet type, online newspapers were the most common workplace (29.2%), followed by freelance journalists (26.2%). The vast majority were currently working as journalists (76.5%), most employed full time (71.2%), and the sample skewed heavily toward experienced professionals, with over 80% having worked in the sector for more than five years. Editors and reporters together accounted for over 80% of the sample. Full sample characteristics are reported in Table C.19.

In this survey, we elicited journalists' beliefs about how using AI-generated images in online articles affects demand for news for that outlet. We did so by leveraging results from Experiment 2. First, we asked respondents about the newsletter sign-up rate in the Control group and then showed it to them as an anchor to calibrate beliefs about baseline demand. We then elicited journalists' beliefs about the newsletter sign-up rate when the article included either a Hard-to-Recognize AI image or an Easy-to-Recognize AI image (presented in random order to mitigate order effects). Next, we elicited beliefs about sign-up rates under the same image conditions when a label was also shown. Finally, we incentivized accuracy using data from Experiment 2: we informed journalists that they would receive an additional $0.50 bonus if their answers exactly matched the Experiment 2 results.

Figure 8a reveals that journalists' beliefs about the effect of AI-generated images on newsletter sign-up rates exhibit a systematic miscalibration and a striking disconnect with the actual experimental results. Journalists believed that approximately 52 out of 100 readers would sign up in the control condition, substantially overestimating the actual sign-up rate of 35 observed in Experiment 2. Beyond this baseline miscalibration, journalists expected AI-generated images to meaningfully depress demand. Anticipated sign-up rates fell to 38.5 for hard-to-recognize AI images and further to 24.4 for easy-to-recognize ones a gap of nearly 14 percentage points suggesting that journalists believe image recognizability plays an important role in shaping reader behavior. They also viewed labeling as consequential: in the hard-to-recognize condition, the expected sign-up rate fell further to 28.8 when a label was added, while beliefs in the easy-to-recognize condition remained largely unchanged at 23.2. Crucially, however, none of these anticipated effects are borne out by the data: Experiment 2 finds no statistically significant difference in demand across any of the image conditions. Journalists thus not only overestimate

baseline engagement, but also overestimate the extent to which readers notice, and respond to, the AI origin of images.[27]

This disconnect between anticipated and realized reputational costs echoes evidence from labor markets that image concerns are a first-order barrier to AI adoption. Almog (2025) finds in a field experiment that workers adopt AI recommendations at lower rates when their reliance on AI is visible to an evaluator, degrading task performance, even though the evaluator is explicitly instructed to reward expected accuracy on the same AI-assisted task and even when reassured about the worker's track record. The operative concern is not that AI lowers output quality but that visible reliance signals an unfavorable type—a lack of confidence in one's own judgment. Our practitioners exhibit the demand-side analogue of this mechanism: journalists anticipate that readers will notice and penalize AI imagery (a gradient of nearly 14 p.p. across recognizability conditions) and accordingly treat AI visuals as reputationally costly, yet our experimental audiences impose no such average penalty. In both settings, the behavior of the AI adopter is governed by an anticipated inference about type rather than by any realized change in quality, and—as in Almog (2025)—reassurance does little to dislodge it.

We further explored journalists' views on using AI to generate images and write articles, as well as their perceptions of how widespread this behaviour is among their colleagues. Figure 8b shows that journalists display strong personal support for transparency around AI-generated content: over 90% agreed that AI use should be clearly disclosed in articles, with only a small minority expressing any disagreement. Views on personal AI adoption are more divided, with roughly equal shares reporting having used or never used AI in their work.

Figure 8c shows journalists' second-order beliefs about their peers, which reveal a degree of social projection. The distribution of estimated peer usage of AI is wide and dispersed across the full 0–100 range, with a mean of approximately 50 out of 100 somewhat higher than journalists' own self-reported behaviour would suggest. By contrast, estimated peer support for disclosure is more concentrated toward higher values, yet centres around 64 out of 100 on average, well below the over 90% observed in the sample itself, suggesting that journalists substantially underestimate how widely pro-disclosure norms are shared among their peers.

When asked why they believed journalists use AI in their work, respondents identified a range of motivations, with efficiency and research support emerging as the two most prominent.[28] Over a third (35.7%) pointed to time pressure and tight deadlines, while 31% highlighted AI's role in research, data processing, and information synthesis. Professional concerns and ethical issues were raised by 29.4% of respondents, with some expressing strong opposition to AI

[27] Appendix B.12 formalises the outlet-level cost–benefit calculation implicit in practitioners' forecasts, anchoring parameters to wire-service licensing rates and industry benchmarks for newsletter revenue.

[28] Open-text responses were coded inductively using a two-step procedure: a language model first assigned granular topic labels to each response, then collapsed them into broad thematic categories. Categories are not mutually exclusive; respondents could mention multiple themes. Percentages reflect the share of the 126 respondents who provided a non-empty answer.

use and others calling for clearer disclosure norms. A further 27% cited workflow facilitation and content scaling, framing AI as a tool to meet high output demands. Image generation and writing and editing support were each cited by 22.2% of respondents: the former noted AI as a practical substitute for costly or unavailable stock photography, while the latter covered tasks such as grammar checking, editing, and drafting. Finally, 12.7% pointed to transparency and ethical frameworks, reflecting awareness of emerging disclosure norms and the need for editorial guidelines to govern AI adoption in the newsroom.

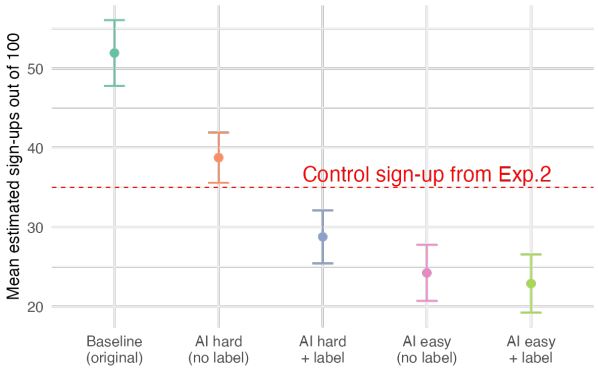
7 Conclusion

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 shows higher engagement; when they do recognise it, engagement is comparably lower. Third, outlet perceptions move in parallel. AI imagery is linked to higher perceived quality, accuracy, and trustworthiness among respondents who think the image is real, and lower for 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 back-of-the-envelope calculation makes this precise: the demand gap between AI and real imagery, currently about half a percentage point and statistically undetectable, widens to 6.2 percentage points under full information and becomes detectable at our sample scale once roughly 80% of audiences correctly identify photorealistic AI images (up from 44% today) while false attributions of authentic photographs fall from 37% to about 13%, a configuration consistent with broad rollout of detection tools, authentication standards for real images, and sustained media-literacy interventions.

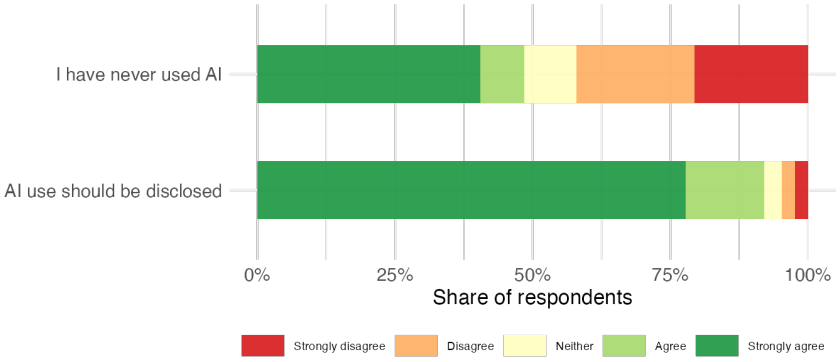
Our initial design is intentionally narrow to isolate this mechanism: one reputable outlet, one article on a politically salient but geographically distant event, compositionally matched imagery, and short-run outcomes in a UK online sample. Nonetheless, core findings replicate across six diverse articles and twelve AI-generated images in Experiment 2: across all article-image pairs, newsletter demand varies by no more than 2.4 p.p. across treatment arms, with none significantly different from control. 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).

Figure 8: Journalists' perceptions

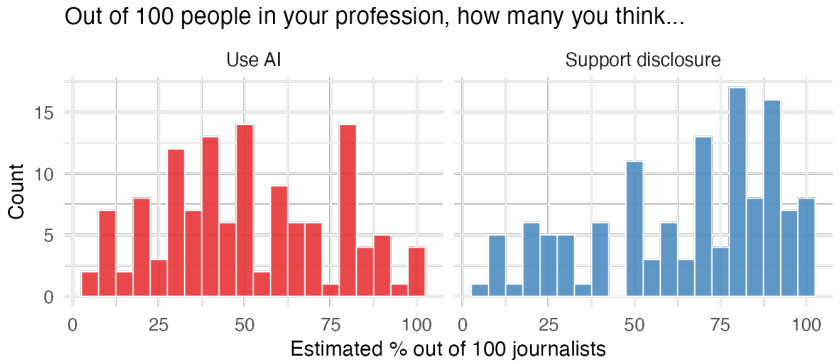
(a) Beliefs about Newsletter Sign-up by Treatment



(b) Attitudes toward Personal Use of AI and its Disclosure



(c) Second-order Beliefs about AI Use and Disclosure Norms among Peers



Notes: Panel (a) displays mean beliefs about newsletter sign-up rates under each image condition, benchmarked against the actual Experiment 2 results. Panel (b) shows the distribution of journalists' personal attitudes toward AI use and disclosure. Panel (c) presents journalists' second-order beliefs about the prevalence of AI use and support for its disclosure among peers. $N = 132$.

Experiment 2 also addresses the key identification challenge in Experiment 1: as recognition of the image source is elicited after the newsletter decision, the heterogeneity by recognition could, in principle, reflect selection rather than a causal belief mechanism. To isolate causality, we randomly vary the photorealism of the AI-generated images. Respondents assigned a less photorealistic AI image are significantly more likely to identify it as synthetic than those assigned a highly photorealistic version, while the depicted scene is held constant across both. Using the exogenous variation in recognition due to treatment assignment, we find that marginally rising the probability of correctly recognizing the image as AI leaves newsletter demand unchanged (2SLS coefficient: $+0.015$, s.e. 0.050, $p > 0.10$), consistent with the aggregate null arising from precisely offsetting belief-driven responses rather than from an absence of any behavioural mechanism, while worsening perceived outlet quality by 0.345 standard deviations ($p < 0.001$). These estimates are consistent in sign and magnitude with the OLS heterogeneity in Experiment 1, supporting the interpretation that belief updating about image origin rather than the image itself or its label is the operative mechanism driving both engagement and credibility responses.

Finally, to assess how industry practitioners perceive these dynamics, we surveyed 132 international journalists recruited via Prolific. Eliciting their beliefs about how AI imagery affects audience demand (anchored to the baseline newsletter sign-up rate from Experiment 2). We find that journalists are systematically miscalibrated in two directions. First, they substantially overestimate baseline newsletter demand, believing approximately 52 out of 100 readers would sign up compared to the actual rate of 35 observed in Experiment 2. Second, they anticipate that AI-generated images meaningfully depress demand expecting sign-up rates of 38.5 for hard-to-recognize AI images and 24.4 for easy-to-recognize ones a gradient of nearly 14 p.p. that finds no support in the data. At the same time, journalists display near-unanimous personal support for disclosure (91.1% in favour), yet underestimate their peers' support (around 50 out of 100) while believing AI adoption among colleagues is considerably more widespread than their own self-reported behaviour implies. This disconnect between practitioner beliefs and empirical responses runs in both directions: journalists overestimate how much readers notice and penalise AI imagery, while simultaneously underestimating how widespread pro-disclosure norms already are among their peers, suggesting that the reputational risks of AI visual adoption may not be fully internalised in current newsroom decision-making.

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. We translate this into an explicit breakeven condition: the per-image cost saving from AI imagery exceeds the engagement loss for micro-outlets even under full information, but because photorealistic AI is already attributed to AI slightly more often than authentic photographs, adoption is unprofitable *today* for any outlet with more than roughly 2,200 effective readers per article, and for smaller outlets the threshold is crossed with only a modest improvement in AI literacy.

References

Alabrese, Eleonora, Francesco Capozza, and Prashant Garg, "Politicized Scientists: Credibility Cost of Political Expression on Twitter," NICEP Working Paper 2025-09, Nottingham Interdisciplinary Centre for Economic and Political Research 2025. Also CESifo Working Paper 11254 and CAGE Working Paper 735.

Almog, David, "Barriers to AI adoption: Image concerns at work," *arXiv preprint arXiv:2511.18582*, 2025.

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.

Associated Press, "AP Images: Editorial Licensing," https://www.apimages.com 2024. Accessed February 2026.

Bammey, Quentin, "Synthbuster: Towards detection of diffusion model generated images," *IEEE Open Journal of Signal Processing*, 2024, *5*, 1–9.

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.

Bastounis, Anastasios, John Buckell, Jamie Hartmann-Boyce, Brian Cook, Sarah King, Christina Potter, Filippo Bianchi, Mike Rayner, and Susan A. Jebb, "The Impact of Environmental Sustainability Labels on Willingness-to-Pay for Foods: A Systematic Review and Meta-Analysis of Discrete Choice Experiments," *Nutrients*, 2021, *13* (8), 2677.

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

Berger, Lara Marie, Anna Kerkhof, and Nikola Noske, "Improving science literacy in the newsroom: Experimental evidence," *PNAS Nexus*, 2026, *5* (4), pgag099.

Bleich, Sara N., Christina D. Economos, Marie L. Spiker, Kristin A. Vercammen, Eric M. VanEpps, Jason P. Block, Brian Elbel, Mary Story, and Christina A. Roberto, "A Systematic Review of Calorie Labeling and Modified Calorie Labeling Interventions: Impact on Consumer and Restaurant Behavior," *Obesity*, 2017, *25* (12), 2018–2044.

Brewer, Noel T., Marissa G. Hall, Seth M. Noar, Humberto Parada, Jeffrey Yin, Linda D. Cameron, and Kurt M. Ribisl, "Effect of Graphic Warning Labels on Cigarette Packs on US Smokers' Cognitions and Smoking Behavior After 3 Months: A Randomized Clinical Trial," *JAMA Network Open*, 2021, *4* (8), e2120245.

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.

Chen, Baoying, Jishen Zeng, Jianquan Yang, and Rui Yang, "DRCT: Diffusion Reconstruction Contrastive Training towards Universal Detection of Diffusion Generated Images.," in "ICML," Vol. 2 2024, p. 6.

Cho, Tony, Allen H. Huang, Joseph Pacelli, and Kristina Rennekamp, "Generative AI and Investor Processing of Financial Media," Working Paper, The Hong Kong University of Science and Technology 2026.

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.

Dranove, David and Ginger Zhe Jin, "Quality Disclosure and Certification: Theory and Practice," *Journal of Economic Literature*, December 2010, *48* (4), 935–963.

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.

Getty Images, "Editorial Image Licensing," https://www.gettyimages.co.uk/editorial-images 2024. Accessed February 2026.

Goldberg, Samuel and H. Tai Lam, "Generative AI in Equilibrium: Evidence from a Creative Goods Marketplace," Working Paper 4265, Stanford Graduate School of Business May 2025. SSRN 5152649.

Hainmueller, Jens, Michael J. Hiscox, and Sandra Sequeira, "Consumer Demand for the Fair Trade Label: Evidence from a Multistore Field Experiment," *Review of Economics and Statistics*, 2015, *97* (2), 242–256.

Hammond, David, "Health warning messages on tobacco products: a review," *Tobacco Control*, 2011, *20* (5), 327–337.

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

Interactive Advertising Bureau UK, "Digital Adspend Report 2023," Technical Report, IAB UK, London 2024.

Kamenica, Emir and Matthew Gentzkow, "Bayesian Persuasion," *American Economic Review*, October 2011, *101* (6), 2590–2615.

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.

Li, Fan and Ya Yang, "Impact of Artificial Intelligence–Generated Content Labels on Perceived Accuracy, Message Credibility, and Sharing Intentions for Misinformation: Web-Based, Randomized, Controlled Experiment," *JMIR Formative Research*, 2024, *8*, e60024.

Li, Ouxiang, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng, "Improving synthetic image detection towards generalization: An image transformation perspective," in "Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 1" 2025, pp. 2405–2414.

Litmus Software, Inc., "2024 State of Email Report," Technical Report, Litmus, Boston, MA 2024.

Martel, Cameron and David G. Rand, "Misinformation warning labels are widely effective: A review of warning effects and their moderating features," *Current Opinion in Psychology*, 2023, *54*, 101745.

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

Muller, Laurent, Anne Lacroix, and Jayson L. Lusk, "Effects of front-of-pack labels on the nutritional quality of supermarket food purchases: Evidence from a large-scale randomized controlled trial," *Journal of the Academy of Marketing Science*, 2021, *49* (4), 700–720.

Nahas, Kamal, "Illustrators Call Out Journals and News Sites for Using AI Art," Nature Index, Nature June 2025.

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

Newman, Nic, Richard Fletcher, Anne Schulz, Simge Andi, and Rasmus Kleis Nielsen, "Reuters Institute Digital News Report 2020," Technical Report, Reuters Institute for the Study of Journalism, University of Oxford 2020.

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, Jeongsoo and Andrew Owens, "Community forensics: Using thousands of generators to train fake image detectors," in "2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)" IEEE 2025, pp. 8245–8257.

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.

Pennycook, Gordon, Adam Bear, Evan T. Collins, and David G. Rand, "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings," *Management Science*, 2020, *66* (11), 4944–4957.

__ **and David G. Rand**, "Shifting attention to accuracy can reduce misinformation online," *Nature*, 2021, *592*, 590–595.

Reimers, Imke and Joel Waldfogel, "AI and the Quantity and Usage of Creative Products: Have LLMs Boosted Creation of Valuable Books?," NBER Working Paper 34777, National Bureau of Economic Research, Cambridge, MA January 2026.

Ren, Simiao, Yuchen Zhou, Xingyu Shen, Kidus Zewde, Tommy Duong, George Huang, En Wei, Jiayu Xue et al., "How well are open sourced AI-generated image detection models out-of-the-box: A comprehensive benchmark study," *arXiv preprint arXiv:2602.07814*, 2026.

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.

Sampalean, Nicoleta, Zoltán Bakucs, Imre Fertő, and Matthew Gorton, "How does origin labelling on food packaging influence consumer product evaluation and choices? A systematic literature review," *Food Policy*, 2023, *118*, 102478.

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.

Stiglitz, Joseph E. and Màxim Ventura-Bolet, "The Impact of AI and Digital Platforms on the Information Ecosystem," NBER Working Paper 34318, National Bureau of Economic Research, Cambridge, MA October 2025. Forthcoming in *The Economics of Transformative AI*, ed. Agrawal, Brynjolfsson and Korinek.

The Conversation, "About The Conversation UK: Our Numbers," https://theconversation.com/uk/about 2024. Accessed February 2026.

Toff, Benjamin and Felix M. Simon, ""Or They Could Just Not Use It?": The Dilemma of AI Disclosure for Audience Trust in News," *The International Journal of Press/Politics*, 2024. Advance online publication.

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.

Zhang, Zirui, Yinbo Yu, Donghai Guan, Chunwei Tian, Daoqiang Zhang, and Qi Zhu, "A benchmark dataset for MLLM-generated image detection: GPT Image2 and Nano Banana2," *arXiv preprint arXiv:2608.01258*, 2026.

Zhao, Hangcheng and Ron Berman, "Strategic response of news publishers to generative AI," *arXiv preprint arXiv:2512.24968*, 2025.

__ **and** __ , "Strategic Response of News Publishers to Generative AI," Working Paper arXiv:2512.24968, arXiv 2026.

Zhong, Nan, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang, "Patchcraft: Exploring texture patch for efficient ai-generated image detection," *arXiv preprint arXiv:2311.12397*, 2023.

Zhu, Kai, "Selective Participation in the AI Data Commons," Working Paper, Bocconi University 2026.

For online publication only:

AI Images, Trust and News Demand

Maja Adena Eleonora Alabrese
Francesco Capozza Isabelle Leader Jaime Marques-Pereira

A AI Image Detection

A.1 AI Image Detectors

To identify AI-generated images, we employ four open-source image-forensic models selected from the benchmark study of Ren et al. (2026), which evaluates the out-of-the-box performance of publicly available AI-image detectors on unseen image generators. The selected models rank among the highest-performing open-source detectors in terms of cross-generator generalization.

Community-Forensics. Community-Forensics (Park and Owens, 2025) is a data-centric detector trained on approximately 2.7 million images generated by 4,803 distinct diffusion models. The detector emphasizes diversity in its training data, demonstrating that exposure to a wide range of image generators substantially improves generalization to previously unseen generative models. In the benchmark of Ren et al. (2026), Community-Forensics is the highest-performing open-source detector. We use the official implementation provided by the authors.

SAFE. SAFE (Synthetic Image Detection via Augmentation-Free Enhancement) (Li et al., 2025) improves generalization by modifying the detector's training procedure rather than introducing a new network architecture. During training, standard image resizing is replaced with cropping, while colour jittering, rotation, and patch-based random masking are used to encourage the detector to learn generator-invariant image characteristics. SAFE ranks second among open-source detectors in the benchmark of Ren et al. (2026). We use the official implementation released by the authors.

DRCT-2.0. DRCT (Diffusion Reconstruction Contrastive Training) (Chen et al., 2024) detects synthetic images by first reconstructing images through a diffusion process before applying contrastive learning to distinguish real images from both generated and reconstructed images. The DRCT-2.0 model is trained using images generated by Stable Diffusion version 2.0 and

demonstrates strong generalization across multiple image generators. It ranks fourth among open-source detectors in Ren et al. (2026). We use the official implementation released by the authors.

DRCT-1.4. DRCT-1.4 employs the same architecture and training procedure as DRCT-2.0 but is trained using images generated by Stable Diffusion version 1.4 (Chen et al., 2024). Evaluating both checkpoints allows us to assess whether the underlying diffusion model used during detector training affects performance on contemporary generative models. DRCT-1.4 ranks fifth among open-source detectors in the benchmark of Ren et al. (2026). We again use the authors' official implementation and pretrained model.

PatchCraft (Zhong et al., 2023), which ranks third in the benchmark of Ren et al. (2026), is not included in our analysis because, at the time of implementation, its source code was not distributed through a maintained public repository and lacked sufficient documentation to support reliable large-scale deployment.

A.2 Agreement Across AI Image Detectors

Although all four detectors are designed to distinguish AI-generated images from human-produced images, they differ in their training data, model architectures, and optimisation objectives. Consequently, their classifications need not coincide. To assess the extent of agreement, Table A.1 reports the weighted proportion of downloaded image occurrences classified as AI-generated by each detector individually (diagonal entries) and jointly by every detector pair (off-diagonal entries).

Across both datasets, substantial differences emerge in the propensity of individual detectors to classify images as AI-generated. Community-Forensics and SAFE are comparatively conservative, whereas the two DRCT models classify roughly three quarters of image occurrences as AI-generated. Despite these differences in overall classification rates, the two DRCT models exhibit high pairwise agreement, jointly classifying more than 70% of weighted image occurrences as AI-generated in both datasets. By contrast, the proportion of images jointly classified as AI-generated by Community-Forensics and SAFE, or by either of these detectors and the DRCT models, is substantially lower.

These patterns indicate that estimates of AI-image prevalence can depend substantially on the detector used when classifications are based on a single model. However, agreement between detectors does not necessarily imply classification accuracy, since detectors may share systematic errors or respond similarly to image characteristics unrelated to AI generation. Rather than selecting a classification rule based solely on agreement within the newspaper corpus, we therefore evaluate alternative ensemble voting rules using an internal validation sample of newspaper images and two external validation datasets.

Table A.1: Agreement across AI-image detectors

Panel A: CC-News

	Community-Forensics	SAFE	DRCT 2.0	DRCT 1.4
Community-Forensics	4.34	1.63	3.66	3.52
SAFE	1.63	3.81	3.29	3.06
DRCT 2.0	3.66	3.29	77.89	69.37
DRCT 1.4	3.52	3.06	69.37	74.38

Panel B: GDELT VGKG

	Community-Forensics	SAFE	DRCT 2.0	DRCT 1.4
Community-Forensics	2.17	0.11	1.82	1.66
SAFE	0.11	1.51	1.28	1.20
DRCT 2.0	1.82	1.28	82.35	72.70
DRCT 1.4	1.66	1.20	72.70	75.51

Panel C: Wayback Machine

	Community-Forensics	SAFE	DRCT 2.0	DRCT 1.4
Community-Forensics	4.30	0.29	3.50	3.57
SAFE	0.29	2.87	2.17	1.93
DRCT 2.0	3.50	2.17	76.36	69.49
DRCT 1.4	3.57	1.93	69.49	75.84

Notes: This table reports unweighted pairwise overlap in AI-image classifications across four detectors for CC-News, GDELT VGKG, and Wayback Machine. CC-News and GDELT VGKG contain sampled news articles, whereas Wayback Machine contains archived newspaper homepage snapshots. For CC-News, statistics are restricted to downloaded image occurrences associated with a nonempty figure caption. For GDELT VGKG and Wayback Machine, statistics use all downloaded image occurrences. Diagonal entries report the unweighted percentage of image occurrences for which the corresponding detector assigns a score greater than or equal to 0.50. Off-diagonal entries report the unweighted percentage of image occurrences for which both corresponding detectors assign scores greater than or equal to 0.50. Each percentage is calculated among image occurrences with finite scores for the corresponding detector or detector pair. The percentages measure joint positive classifications rather than overall agreement, which would also include joint negative classifications. Image occurrences include repeated appearances of the same image. For Wayback Machine, additional homepage snapshots have been sampled, but their images are still being downloaded and scored.

A.3 Validation of AI Image Detectors

We assess the performance of the four AI-image detectors using an internal validation sample of newspaper images with provenance information in their captions and two external validation datasets, Synthbuster (Bammey, 2024, a dataset covering generators up to 2023) and MLLM

(Zhang et al., 2026, a dataset covering more recent generators).[1] This analysis complements the detector agreement results in Section A.2 and informs the selection of the classification rule used in the main analysis. The external datasets allow us to assess whether classification performance extends beyond the newspaper images used for internal validation, including images produced by different generations of AI image-generation systems.

To construct the internal validation sample, we begin with CC-News images that were successfully downloaded and have a nonempty caption and valid scores from all four detectors. We use captions written in English or successfully translated into English. To identify AI-generated images, we search for expressions such as "AI-generated" and "AI illustration," as well as references to AI image-generation tools. To identify human-produced images, we use photographer credits, such as "photo by," and photographic copyright information. We exclude observations containing both types of evidence. Lastly, we remove repeated images and images with conflicting provenance information across captions.

This procedure yields 159 distinct images with AI provenance. We retain all of them and select 159 images with human provenance for comparison. We match the two groups by country, news outlet, and year whenever possible, using less restrictive criteria when exact matches are not available. The final internal validation sample thus contains 318 images, with equal numbers in each group.

We complement this internal sample with two external validation datasets, Synthbuster (Bammey, 2024) and MLLM (Zhang et al., 2026). For each dataset, we evaluate the detectors on a balanced main sample of 159 AI-generated and 159 human-produced images. Using these external datasets allows us to examine performance on images generated outside the newspaper-caption sampling procedure and to assess differences in performance across image-generation settings.

We evaluate each detector using a score threshold of 0.50 and consider 28 ensemble voting rules constructed from combinations of two, three, or four detectors. An ensemble classifies an image as AI-generated when the number of positive detector votes meets or exceeds its specified threshold. We evaluate performance using sensitivity, specificity, the F1 score, and the Matthews correlation coefficient (MCC). Sensitivity measures the proportion of AI-generated images correctly identified, whereas specificity measures the proportion of human-produced images correctly classified. MCC summarizes classification performance using all four entries of the confusion matrix and is particularly useful when assessing the trade-off between false positives and false negatives.

We also quantify uncertainty in the estimated performance measures. For the internal validation sample, we report 95% bootstrap confidence intervals. For the external datasets, we report both 95% bootstrap confidence intervals and 95% percentile intervals calculated

[1] We selected the external validation datasets after reviewing the original publications of the four detectors to minimize potential overlap between the images used for evaluation and those used to train the detectors.

across alternative stratified samples. The latter describe the sensitivity of the results to the composition of the selected external samples rather than conventional population confidence intervals. Table A.2 reports the performance of the four individual detectors, while Table A.3 reports the four highest-ranked ensemble rules.

The internal validation results reveal substantial differences across individual detectors. Community-Forensics identifies 22.01% of AI-provenance images and correctly classifies 98.11% of human-provenance images. SAFE identifies only 6.29% of AI-provenance images but produces no false positives. By contrast, DRCT 2.0 and DRCT 1.4 identify 86.16% and 78.62% of AI-provenance images, respectively. However, they also incorrectly classify many human-provenance images as AI-generated, with specificities of only 22.64% and 29.56%. These results illustrate the substantial trade-off between sensitivity and specificity when relying on individual detectors.

The external validation results provide an additional perspective on this trade-off. Performance differs across the validation datasets, indicating that conclusions drawn from the internal newspaper sample alone may not generalize to images produced by other generation systems. In particular, a detector or ensemble that achieves high specificity in one dataset may exhibit different sensitivity or specificity in another. We therefore consider performance across all three datasets when selecting an ensemble, rather than optimizing exclusively for the internal sample.

Specifically, we rank the 28 ensemble rules according to their equally weighted mean MCC across the internal validation sample, Synthbuster, and MLLM. This approach favors ensembles with strong average classification performance while also considering their weakest observed performance across validation settings. We prioritize MCC because, unlike the F1 score, it explicitly accounts for true negatives. This is particularly important in our setting, where we aim not only to correctly identify AI-generated images but also to minimize the misclassification of genuine newspaper images as AI-generated.

Table A.3 presents the four highest-ranked ensembles under this criterion. The leading rule combines Community-Forensics, SAFE, and DRCT 2.0, classifying an image as AI-generated when at least two of the three detectors vote positive. The remaining leading rules include a three-of-four-detector voting rule, a majority vote combining Community-Forensics, SAFE, and DRCT 1.4, and a rule requiring agreement between Community-Forensics and DRCT 2.0. The comparison illustrates that relatively restrictive voting thresholds can improve specificity, although this generally comes at the cost of lower sensitivity.

This validation has several limitations. First, image provenance in the internal sample is inferred from captions rather than independently verified by human coders. Captions may contain incorrect or incomplete information, and images explicitly identified as AI-generated may differ from AI-generated images without such labels. Second, although the external datasets broaden the range of images considered and allow us to assess detector robustness across

different generations of AI-generated images, they do not necessarily represent the distribution of AI-generated and human-produced imagery in newspaper archives. Third, all three validation samples contain equal numbers of AI-generated and human-produced images by construction. This class balance is unlikely to reflect the distribution of images in the newspaper corpus, where AI-generated images are likely to be less prevalent. Because some performance metrics, particularly precision and the F1 score, depend on class prevalence, their values in our balanced validation samples may not accurately reflect detector performance in the full newspaper corpus.

The selected rule prioritizes reliable identification of AI-generated images while retaining a relatively low sensitivity in the internal validation sample. It is therefore likely to miss some AI-generated images in the newspaper corpus. Our main results should consequently be interpreted as the share of images classified as AI-generated by the selected ensemble, rather than as a direct estimate of the true prevalence of AI-generated imagery in newspapers.

A.4 Prevalence of AI-Generated Images by Country

Table A.4 reports the estimated prevalence of AI-generated images classified as AI-generated by the selected ensemble separately for each of the seventeen countries in our sampling frame, for both the CC-News and GDELT VGKG samples. The cross-country patterns are discussed in Section 2.

A.5 Prevalence of AI-Generated Images by Topic

Table A.5 reports the estimated prevalence of AI-generated images classified as AI-generated by the selected ensemble separately by article topic, for both the CC-News and GDELT VGKG samples. Topics are the top-level categories of the IPTC Media Topics taxonomy; the assignment procedure is described in Appendix A.8. The patterns are discussed in Section 2.

Table A.6 reports the corresponding prevalence of AI disclosures in image captions by article topic, using the disclosure definitions of Appendix A.7.

A.6 AI-Image Prevalence and Outlet Readership

Figure A.1 relates outlet-level AI-image prevalence to weekly online readership, as discussed in Section 2. Panel (a) plots the raw relationship. Panel (b) presents a complementary specification in which both weekly online readership and the share of images classified as AI-generated are residualized with respect to year fixed effects, so that the relationship is net of common changes in readership and AI adoption over time.

Table A.2: Validation performance of individual AI-image detectors

Metric	CommunityForensics	SAFE	DRCT20	DRCT14
Internal validation				
MCC	0.310 [0.230, 0.383]	0.180 [0.113, 0.230]	0.114 [0.008, 0.213]	0.094 [-0.014, 0.203]
F1-score	0.355 [0.269, 0.439]	0.118 [0.049, 0.183]	0.654 [0.621, 0.684]	0.631 [0.591, 0.670]
Sensitivity	0.220 [0.157, 0.283]	0.063 [0.025, 0.101]	0.862 [0.805, 0.918]	0.786 [0.717, 0.849]
Specificity	0.981 [0.962, 1.000]	1.000 [1.000, 1.000]	0.226 [0.164, 0.296]	0.296 [0.226, 0.371]
Synthbuster				
MCC	0.649 B: [0.568, 0.726] S: [0.550, 0.693]	-0.051 B: [-0.162, 0.060] S: [-0.168, -0.015]	0.723 B: [0.651, 0.792] S: [0.643, 0.775]	0.601 B: [0.516, 0.682] S: [0.557, 0.701]
F1-score	0.819 B: [0.775, 0.860] S: [0.764, 0.841]	0.583 B: [0.540, 0.623] S: [0.556, 0.597]	0.867 B: [0.834, 0.899] S: [0.832, 0.891]	0.810 B: [0.769, 0.847] S: [0.789, 0.856]
Sensitivity	0.799 B: [0.736, 0.855] S: [0.723, 0.830]	0.730 B: [0.660, 0.799] S: [0.698, 0.761]	0.925 B: [0.881, 0.962] S: [0.918, 0.981]	0.855 B: [0.799, 0.906] S: [0.830, 0.925]
Specificity	0.849 B: [0.792, 0.906] S: [0.792, 0.893]	0.226 B: [0.164, 0.296] S: [0.132, 0.245]	0.792 B: [0.730, 0.855] S: [0.679, 0.805]	0.742 B: [0.673, 0.805] S: [0.686, 0.805]
MLLM 2026				
MCC	0.451 B: [0.397, 0.497] S: [0.385, 0.497]	0.795 B: [0.748, 0.841] S: [0.746, 0.848]	0.118 B: [0.014, 0.212] S: [0.080, 0.282]	0.058 B: [-0.040, 0.154] S: [0.057, 0.249]
F1-score	0.519 B: [0.456, 0.575] S: [0.444, 0.578]	0.878 B: [0.846, 0.908] S: [0.848, 0.915]	0.632 B: [0.594, 0.668] S: [0.607, 0.688]	0.579 B: [0.534, 0.621] S: [0.556, 0.651]
Sensitivity	0.352 B: [0.302, 0.409] S: [0.289, 0.409]	0.792 B: [0.742, 0.837] S: [0.755, 0.862]	0.767 B: [0.711, 0.824] S: [0.711, 0.830]	0.648 B: [0.585, 0.711] S: [0.585, 0.717]
Specificity	0.994 B: [0.981, 1.000] S: [0.975, 1.000]	0.987 B: [0.969, 1.000] S: [0.956, 0.994]	0.340 B: [0.270, 0.409] S: [0.333, 0.478]	0.409 B: [0.340, 0.478] S: [0.428, 0.572]

Notes: This table reports the validation performance of four individual AI-generated image detectors. Sensitivity is the true-positive rate; specificity is the true-negative rate; F1 is the harmonic mean of precision and sensitivity; MCC is the Matthews correlation coefficient. Point estimates are based on the main validation samples. Bracketed values are 95% percentile intervals. Internal validation reports bootstrap confidence intervals. External validation reports bootstrap confidence intervals (B) and alternative-sample percentile intervals (S).

Table A.3: Validation performance of ensembles of AI-image detectors

Metric	E1	E2	E3	E4
Internal validation				
MCC	0.316 [0.237, 0.388]	0.298 [0.217, 0.370]	0.310 [0.230, 0.383]	0.291 [0.214, 0.364]
F1-score	0.364 [0.278, 0.444]	0.338 [0.251, 0.422]	0.355 [0.269, 0.437]	0.330 [0.247, 0.412]
Sensitivity	0.226 [0.164, 0.289]	0.208 [0.145, 0.270]	0.220 [0.157, 0.283]	0.201 [0.145, 0.264]
Specificity	0.981 [0.962, 1.000]	0.981 [0.962, 1.000]	0.981 [0.962, 1.000]	0.981 [0.962, 1.000]
Synthbuster				
MCC	0.655 B: [0.580, 0.731] S: [0.524, 0.669]	0.750 B: [0.680, 0.818] S: [0.642, 0.775]	0.618 B: [0.542, 0.698] S: [0.507, 0.655]	0.745 B: [0.677, 0.810] S: [0.674, 0.786]
F1-score	0.836 B: [0.801, 0.871] S: [0.779, 0.843]	0.870 B: [0.831, 0.907] S: [0.820, 0.888]	0.820 B: [0.784, 0.856] S: [0.770, 0.837]	0.843 B: [0.792, 0.887] S: [0.794, 0.874]
Sensitivity	0.899 B: [0.855, 0.943] S: [0.855, 0.931]	0.843 B: [0.786, 0.899] S: [0.805, 0.899]	0.887 B: [0.836, 0.931] S: [0.830, 0.918]	0.742 B: [0.673, 0.805] S: [0.679, 0.792]
Specificity	0.748 B: [0.679, 0.818] S: [0.622, 0.755]	0.906 B: [0.862, 0.950] S: [0.805, 0.906]	0.723 B: [0.654, 0.792] S: [0.635, 0.761]	0.981 B: [0.956, 1.000] S: [0.950, 0.994]
MLLM 2026				
MCC	0.747 B: [0.695, 0.803] S: [0.672, 0.791]	0.600 B: [0.541, 0.659] S: [0.553, 0.680]	0.680 B: [0.624, 0.737] S: [0.611, 0.729]	0.360 B: [0.305, 0.409] S: [0.307, 0.424]
F1-score	0.842 B: [0.800, 0.881] S: [0.788, 0.876]	0.707 B: [0.647, 0.762] S: [0.672, 0.783]	0.785 B: [0.736, 0.830] S: [0.728, 0.827]	0.392 B: [0.326, 0.454] S: [0.325, 0.472]
Sensitivity	0.736 B: [0.679, 0.799] S: [0.667, 0.792]	0.553 B: [0.484, 0.623] S: [0.516, 0.654]	0.654 B: [0.591, 0.717] S: [0.585, 0.717]	0.245 B: [0.195, 0.296] S: [0.195, 0.314]
Specificity	0.987 B: [0.969, 1.000] S: [0.956, 1.000]	0.987 B: [0.969, 1.000] S: [0.962, 1.000]	0.987 B: [0.969, 1.000] S: [0.956, 1.000]	0.994 B: [0.981, 1.000] S: [0.981, 1.000]

Notes: This table reports the four best ensemble voting rules selected from 28 candidates by their equally weighted mean Matthews correlation coefficient (MCC) across validation datasets. Sensitivity is the true-positive rate; specificity is the true-negative rate; F1 is the harmonic mean of precision and sensitivity. Point estimates are based on the main validation samples. Bracketed values are 95% percentile intervals. Internal validation reports bootstrap confidence intervals. External validation reports bootstrap confidence intervals (B) and alternative-sample percentile intervals (S). The selected ensembles are: E1: Community-Forensics + SAFE + DRCT 2.0 (rule: at least 2 of 3 detectors must vote positive); E2: 3+ of 4 detectors (rule: at least 3 of 4 detectors must vote positive); E3: Community-Forensics + SAFE + DRCT 1.4 (rule: at least 2 of 3 detectors must vote positive); E4: Community-Forensics + DRCT 2.0 (rule: at least 2 of 2 detectors must vote positive).

Table A.4: AI-image prevalence by country

Country	CC-News	GDELT VGKG	Wayback Machine
AR	4.21	2.50	5.02
	(466)	(418)	(9,171)
AU	2.78	2.42	5.25
	(1,009)	(625)	(9,316)
BR	4.73	4.14	8.32
	(288)	(595)	(15,392)
CA	2.38	2.96	1.86
	(400)	(557)	(1,255)
CO	4.08	2.88	7.91
	(497)	(598)	(12,759)
DE	2.32	2.09	5.64
	(333)	(255)	(17,641)
FR	2.24	2.50	6.73
	(497)	(481)	(6,415)
IN	17.73	4.34	8.67
	(6,219)	(564)	(14,485)
IT	1.67	2.77	4.11
	(168)	(715)	(15,342)
JP	0.77	0.90	1.80
	(36)	(349)	(3,223)
KE	1.55	0.57	3.80
	(208)	(150)	(1,182)
KR	3.84	1.96	3.70
	(187)	(424)	(10,745)
MX	4.39	2.10	8.60
	(538)	(302)	(6,671)
NI	2.65	4.02	5.65
	(224)	(515)	(3,225)
UK	4.88	2.24	5.72
	(1,259)	(1,185)	(3,021)
US	4.90	3.81	3.85
	(416)	(905)	(1,301)
ZA	2.34	3.38	1.72
	(147)	(523)	(40)

Notes: This table reports the weighted share of downloaded image occurrences classified as AI-generated across countries in the CC-News, GDELT VGKG, and Wayback Machine samples. CC-News and GDELT VGKG contain sampled news articles, whereas Wayback Machine contains archived newspaper homepage snapshots. For CC-News, statistics are restricted to downloaded image occurrences associated with a nonempty figure caption. For GDELT VGKG and Wayback Machine, statistics use all downloaded image occurrences. AI-generated images are images classified as AI-generated by an ensemble of AI detectors, selected through a validation exercise comparing different ensemble rules across internal and external image repositories. The first number in each cell is the weighted percentage of image occurrences classified as AI-generated, with the corresponding unweighted count reported in parentheses. Weighted percentages use the inverse sampling probability of the article or homepage snapshot associated with each image occurrence. Image occurrences include repeated appearances of the same image. For Wayback Machine, additional homepage snapshots have been sampled, but their images are still being downloaded and scored.

Table A.5: AI-image prevalence by article topic

IPTC article topic	CC-News	GDELT VGKG
arts, culture, entertainment and media	13.90 (1,811)	2.21 (771)
conflict, war and peace	0.00 (0)	0.00 (0)
crime, law and justice	1.91 (24)	2.53 (67)
disaster, accident and emergency incident	1.61 (3)	0.20 (3)
economy, business and finance	8.94 (640)	2.38 (505)
education	23.54 (245)	2.76 (28)
environment	2.15 (15)	2.66 (23)
health	8.44 (97)	6.07 (110)
human interest	1.55 (48)	1.66 (62)
labour	0.00 (0)	2.51 (5)
lifestyle and leisure	13.81 (1,295)	3.74 (553)
politics	3.38 (195)	2.12 (323)
religion and belief	55.68 (278)	3.18 (1)
science and technology	19.27 (381)	4.68 (271)
society	3.57 (100)	1.84 (81)
sport	6.95 (975)	2.26 (227)
weather	8.55 (12)	4.48 (7)

Notes: This table reports the share of downloaded image occurrences classified as AI-generated by the selected ensemble across IPTC top-level Media Topics in the CC-News and GDELT VGKG samples. Article topics are derived from outlet-native editorial themes identified from article URL paths and harmonised to the IPTC Media Topics taxonomy. The ensemble combines multiple AI-image detectors and was selected based on its performance on a validation sample of newspaper images, achieving the highest balanced accuracy and Matthews correlation coefficient among the classification rules evaluated. For CC-News, statistics are restricted to downloaded image occurrences associated with a nonempty figure caption. For GDELT VGKG, statistics use all downloaded image occurrences. In both samples, statistics are restricted to images with complete detector scores. Each image occurrence inherits the IPTC topic assigned to the article in which it appears. The first number in each cell reports the inverse-probability-weighted percentage of eligible image occurrences within the corresponding article topic that are classified as AI-generated by the selected ensemble. The number in parentheses reports the unweighted count of image occurrences classified as AI-generated within that topic. Weighted percentages use the inverse article-sampling probability associated with each image occurrence.

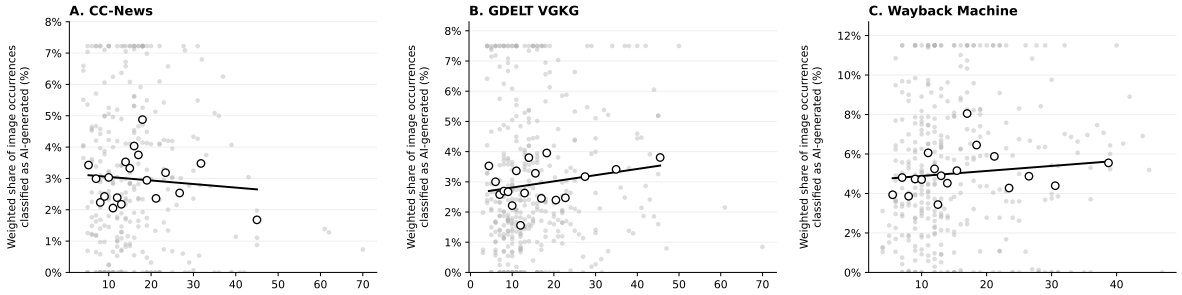
Table A.6: Prevalence of AI disclosures by article topic

	All analyzable captions		Images deemed AI by ≥3 detectors	
IPTC article topic	Strict AI-image	Any AI	Strict AI-image	Any AI
arts, culture, entertainment and media	0.08 (11)	0.08 (11)	0.00 (0)	0.00 (0)
conflict, war and peace	0.00 (0)	0.00 (0)	—	—
crime, law and justice	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
disaster, accident and emergency incident	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
economy, business and finance	0.20 (15)	0.20 (15)	0.16 (1)	0.16 (1)
education	0.69 (6)	0.69 (6)	0.57 (1)	0.57 (1)
environment	0.78 (1)	0.78 (1)	0.00 (0)	0.00 (0)
health	0.32 (3)	0.32 (3)	0.01 (1)	0.01 (1)
human interest	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
labour	0.00 (0)	0.00 (0)	—	—
lifestyle and leisure	0.31 (32)	0.31 (32)	0.89 (11)	0.89 (11)
politics	0.01 (3)	0.01 (3)	0.00 (0)	0.00 (0)
religion and belief	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
science and technology	0.29 (8)	0.29 (8)	0.66 (4)	0.66 (4)
society	0.20 (4)	0.20 (4)	1.69 (1)	1.69 (1)
sport	0.05 (6)	0.05 (6)	0.00 (0)	0.00 (0)
weather	0.42 (3)	0.42 (3)	0.00 (0)	0.00 (0)

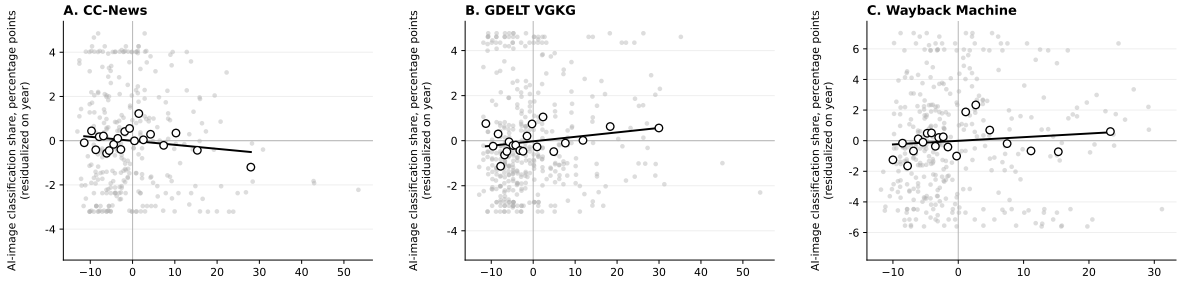
Notes: This table reports AI-disclosure prevalence by IPTC top-level Media Topic among captions associated with downloaded image occurrences in the CC-News sample. Article topics are derived from outlet-native editorial themes identified from article URL paths and harmonised to the IPTC Media Topics taxonomy. Statistics are restricted to exact figure-level image occurrences with a nonempty caption, nonmissing scores from all four AI-image detectors, and captions eligible for English-language disclosure analysis. Each image occurrence inherits the IPTC topic assigned to the article in which it appears. The first two columns report disclosure prevalence among all such analyzable captions, while the final two columns restrict the denominator to captions associated with images classified as AI-generated by at least three of the four detectors. Strict AI-image disclosures identify captions indicating that an image was AI-generated or naming an AI image-generation model. Any AI disclosure additionally includes AI editing disclosures. The first number in each cell is the weighted percentage of eligible image occurrences containing the indicated disclosure; the unweighted number of disclosure-positive image occurrences is reported in parentheses. A value of 0.00 with a count of zero indicates that eligible observations exist for that topic but no disclosures of the indicated type were detected. A dash indicates that no eligible observations exist for the corresponding denominator. Weighted percentages use the inverse article-sampling probability associated with each image occurrence.

Figure A.1: AI-image prevalence and outlet readership

(a) Raw relationship



(b) Controlling for year



Notes: This figure relates the share of downloaded image occurrences classified as AI-generated to weekly online readership, as reported in the Reuters Institute *Digital News Report.* The unit of observation is an outlet–year. AI-image classification is based on the selected ensemble of AI-image detectors, chosen because it achieved the highest mean Matthews correlation coefficient across an internal validation sample of newspaper images and two external validation datasets (Appendix A.3). For CC-News, estimates are restricted to downloaded image occurrences associated with a nonempty figure caption. For GDELT VGKG, estimates use all downloaded image occurrences. Both samples require complete detector scores. Readership observations are drawn from online readership files and matched by country, outlet domain, and year. The share classified as AI-generated is weighted by the inverse article-sampling probability within each outlet–year and winsorized from above at the dataset-specific 90th percentile. Panel (a) reports the raw relationship, while Panel (b) removes year fixed effects from both readership and the winsorized classification share. Light-grey points represent outlet–year observations, hollow circles represent means within 20 readership bins, and solid black lines are linear fits through the binned means. Each outlet–year receives equal weight in the binscatter.

Table A.7: News outlets included: United Kingdom and Australia

Country	News outlets (URL)
United Kingdom	BBC News (https://www.bbc.com/news); Channel 4 News (https://www.channel4.com/news); Channel 5 News (https://www.5news.co.uk); Daily Mail (https://www.dailymail.co.uk); Daily Mirror (https://www.mirror.co.uk); GB News (https://www.gbnews.com); ITV News (https://www.itv.com/news); Metro (https://metro.co.uk); Sky News (https://news.sky.com); The Guardian (https://www.theguardian.com); The Independent (https://www.independent.co.uk); The Sun (https://www.thesun.co.uk); The Telegraph (https://www.telegraph.co.uk); The Times (https://www.thetimes.com); The i (https://inews.co.uk)
Australia	10 News First (https://10play.com.au/news); 7NEWS (https://7news.com.au); 9News (https://www.9news.com.au); ABC News (https://www.abc.net.au/news); Australian Financial Review (https://www.afr.com); Guardian Australia (https://www.theguardian.com/au); Herald Sun (https://www.heraldsun.com.au); MailOnline Australia (https://www.dailymail.co.uk/auhome/index.html); Nine (https://www.nine.com.au); Prime7 (https://www.prime7.com.au); SBS News (https://www.sbs.com.au/news); Sky News Australia (https://www.skynews.com.au); The Age (https://www.theage.com.au); The Australian (https://www.theaustralian.com.au); The Courier-Mail (https://www.couriermail.com.au); The Daily Telegraph (https://www.dailytelegraph.com.au); The New Daily (https://www.thenewdaily.com.au); The Sydney Morning Herald (https://www.smh.com.au); WIN News (https://www.wintv.com.au); news.com.au (https://www.news.com.au)

Notes: This table lists the news outlets included in the descriptive analysis together with the corresponding URLs used to identify articles in the CC-News archive. Outlets are drawn from the Reuters Digital News Report and harmonized across survey waves to produce a unique set of outlet names. The outlet–URL pairs define the sampling frame used throughout the descriptive analysis.

Table A.8: News outlets included: Germany, France, and Italy

Country	News outlets (URL)
Germany	ARD Tagesschau (https://www.tagesschau.de); BILD (https://www.bild.de); Der Spiegel (https://www.spiegel.de); FOCUS (https://www.focus.de); Frankfurter Allgemeine Zeitung (https://www.faz.net); GMX (https://www.gmx.net); ProSieben Newstime (https://www.prosieben.de/serien/newstime); RTL Aktuell (https://www.rtl.de); Sat.1 Nachrichten (https://www.sat1.de); Stern (https://www.stern.de); Süddeutsche Zeitung (https://www.sueddeutsche.de); WEB.DE (https://web.de); WELT (https://www.welt.de); ZDF heute (https://www.zdf.de/nachrichten); ZEIT ONLINE (https://www.zeit.de); n-tv (https://www.n-tv.de); t-online (https://www.t-online.de)
France	20 Minutes (https://www.20minutes.fr); ARTE (https://www.arte.tv/fr); BFM TV (https://www.bfmtv.com); Brut (https://www.brut.media/fr); CNews (https://www.cnews.fr); Canal+ (https://www.canalplus.com); France 24 (https://www.france24.com/fr); France Télévisions (https://www.france.tv); HuffPost France (https://www.huffingtonpost.fr); Konbini (https://www.konbini.com); L'Internaute (https://www.linternaute.com); L'Équipe (https://www.lequipe.fr); La Voix du Nord (https://www.lavoixdunord.fr); Le Figaro (https://www.lefigaro.fr); Le Monde (https://www.lemonde.fr); Le Parisien (https://www.leparisien.fr); Le Point (https://www.lepoint.fr); M6 Info (https://www.m6info.fr); MSN France (https://www.msn.com/fr-fr/actualite); Mediapart (https://www.mediapart.fr); Ouest-France (https://www.ouest-france.fr); RTL (https://www.rtl.fr); TF1 Info (https://www.tf1info.fr); franceinfo (https://www.francetvinfo.fr)
Italy	ANSA (https://www.ansa.it); Corriere della Sera (https://www.corriere.it); Dritto e Rovescio (https://www.mediasetinfinity.mediaset.it/programmi-tv/drittoerovescio_SE000000000892); Fanpage (https://www.fanpage.it); Fuori dal Coro (https://www.mediasetinfinity.mediaset.it/programmi-tv/fuoridalcoro_SE000000000472); Il Fatto Quotidiano (https://www.ilfattoquotidiano.it); Il Post (https://www.ilpost.it); Il Sole 24 Ore (https://www.ilsole24ore.com); La Repubblica (https://www.repubblica.it); La Stampa (https://www.lastampa.it); Libero (https://www.liberoquotidiano.it); Non è l'Arena (https://www.la7.it/nonelarena); Piazzapulita (https://www.la7.it/piazzapulita); Porta a Porta (https://www.raiplay.it/programmi/portaaporta); Quotidiano.net (https://www.quotidiano.net); RaiNews (https://www.rainews.it); Sky TG24 (https://tg24.sky.it); TG LA7 (https://tg.la7.it); TGCOM24 (https://www.tgcom24.mediaset.it)

Notes: This table lists the news outlets included in the descriptive analysis together with the corresponding URLs used to identify articles in the CC-News archive. Outlets are drawn from the Reuters Digital News Report and harmonized across survey waves to produce a unique set of outlet names. The outlet–URL pairs define the sampling frame used throughout the descriptive analysis.

Table A.9: News outlets included: United States, Canada and Mexico

Country	News outlets (URL)
United States	ABC News (https://abcnews.go.com); BuzzFeed (https://www.buzzfeed.com); CBS News (https://www.cbsnews.com); CNN (https://www.cnn.com); Fox News (https://www.foxnews.com); HuffPost (https://www.huffpost.com); MSN (https://www.msn.com/news); NBC News/MSNBC (https://www.nbcnews.com); NPR (https://www.npr.org); New York Post (https://nypost.com); New York Times (https://www.nytimes.com); Newsmax (https://www.newsmax.com); PBS News (https://www.pbs.org/newshour); The Rush Limbaugh Show (https://www.rushlimbaugh.com); USA Today (https://www.usatoday.com); Wall Street Journal (https://www.wsj.com); Washington Post (https://www.washingtonpost.com); Yahoo News (https://news.yahoo.com)
Canada	BNN Bloomberg (https://www.bnnbloomberg.ca); CBC News (https://www.cbc.ca/news); CP24 (https://www.cp24.com); CTV News (https://www.ctvnews.ca); CityNews (https://www.citynews.ca); Global News (https://globalnews.ca); Journal de Montréal/Journal de Québec (https://www.journaldemontreal.com); L'actualité (https://lactualite.com); La Presse (https://www.lapresse.ca); Le Devoir (https://www.ledevoir.com); Narcity (https://www.narcity.com); National Post (https://nationalpost.com); Noovo Info (https://www.noovo.info); QUB (https://www.qub.ca); Radio-Canada (https://ici.radio-canada.ca/info); TVA Nouvelles (https://www.tvanouvelles.ca); The Globe and Mail (https://www.theglobeandmail.com); Toronto Star (https://www.thestar.com)
Mexico	Animal Político (https://animalpolitico.com); Aristegui Noticias (https://aristeguinoticias.com); Azteca Noticias (https://www.tvazteca.com/aztecanoticias); Canal 22 (https://canal22.org.mx); Canal Once (https://canalonce.mx); El Financiero (https://www.elfinanciero.com.mx); El Heraldo de México (https://heraldodemexico.com.mx); El Sol de México (https://www.elsoldemexico.com.mx); El Universal (https://www.eluniversal.com.mx); Excélsior (https://www.excelsior.com.mx); Imagen Noticias (https://www.imagenradio.com.mx); La Jornada (https://www.jornada.com.mx); Latinus (https://latinus.us); Milenio (https://www.milenio.com); N+ (https://www.nmas.com.mx); N+ FORO (https://www.nmas.com.mx/foro-tv); Radio Fórmula (https://www.radioformula.com.mx); Reforma (https://www.reforma.com); UnoTV (https://www.unotv.com)

Notes: This table lists the news outlets included in the descriptive analysis together with the corresponding URLs used to identify articles in the CC-News archive. Outlets are drawn from the Reuters Digital News Report and harmonized across survey waves to produce a unique set of outlet names. The outlet–URL pairs define the sampling frame used throughout the descriptive analysis.

Table A.10: News outlets included: Brazil, Argentina and Colombia

Country	News outlets (URL)
Brazil	Band (https://www.band.uol.com.br); BandNews (https://bandnewsfm.band.uol.com.br); BandNews TV (https://www.band.uol.com.br/bandnews-fm); CNN Brasil (https://www.cnnbrasil.com.br); Extra (https://extra.globo.com); Folha de S.Paulo (https://www.folha.uol.com.br); Globo (https://www.globo.com); GloboNews (https://g1.globo.com/globonews); ISTOÉ (https://istoe.com.br); Jovem Pan News (https://jovempan.com.br); MSN Brasil (https://www.msn.com/pt-br/noticias); Metrópoles (https://www.metropoles.com); O Estado de S. Paulo (https://www.estadao.com.br); O Globo (https://oglobo.globo.com); Record (https://recordtv.r7.com); Record News (https://noticias.r7.com/record-news); RedeTV! (https://www.redetv.uol.com.br); SBT News (https://www.sbtnews.com.br); TV Globo (https://g1.globo.com); Terra (https://www.terra.com.br); UOL (https://www.uol.com.br); VEJA (https://veja.abril.com.br)
Argentina	A24 (https://www.a24.com); América TV (https://www.americatv.com.ar); C5N (https://www.c5n.com); Cadena 3 (https://www.cadena3.com); Canal 26 (https://www.canal26.com); Clarín (https://www.clarin.com); Crónica (https://www.cronica.com.ar); El Cronista (https://www.cronista.com); El Destape (https://www.eldestapeweb.com); El Nueve (https://www.elnueve.com.ar); El Trece (https://www.eltrecetv.com.ar); Infobae (https://www.infobae.com); LN+ (https://www.lanacion.com.ar/lnmas); LUZU TV (https://www.luzutv.com); La 100 (https://www.lacienradios.com); La Nación (https://www.lanacion.com.ar); La Voz (https://www.lavoz.com.ar); Minuto Uno (https://www.minutouno.com); Olé (https://www.ole.com.ar); Primicias Ya (https://www.primiciasya.com); Página/12 (https://www.pagina12.com.ar); Radio Mitre (https://radiomitre.cienradios.com); TN (https://tn.com.ar); TV Pública (https://www.tvpublica.com.ar); Telefe Noticias (https://noticias.telefe.com); Ámbito (https://www.ambito.com)
Colombia	Blu Radio (https://www.bluradio.com); CableNoticias (https://cablenoticias.tv); Caracol Radio (https://caracol.com.co); CityTV (https://citytv.eltiempo.com); El Espectador (https://www.elespectador.com); El País (https://www.elpais.com.co); El Tiempo (https://www.eltiempo.com); La FM (https://www.lafm.com.co); La Silla Vacía (https://www.lasillavacia.com); Las2Orillas (https://www.las2orillas.co); Minuto30 (https://www.minuto30.com); NTN24 (https://www.ntn24.com); Noticentro 1 CM& (https://canal1.com.co/noticias); Noticias Caracol (https://www.noticiascaracol.com); Noticias RCN (https://www.noticiasrcn.com); Noticias Uno (https://www.noticiasuno.com); Pulzo (https://www.pulzo.com); Q'Hubo (https://www.qhubo.com); RCN Radio (https://www.rcnradio.com); Radio Nacional de Colombia (https://www.radionacional.co); Semana (https://www.semana.com); Señal Colombia (https://www.senalcolombia.tv); W Radio (https://www.wradio.com.co)

Notes: This table lists the news outlets included in the descriptive analysis together with the corresponding URLs used to identify articles in the CC-News archive. Outlets are drawn from the Reuters Digital News Report and harmonized across survey waves to produce a unique set of outlet names. The outlet–URL pairs define the sampling frame used throughout the descriptive analysis.

Table A.11: News outlets included: Japan, India and South Korea

Country	News outlets (URL)
Japan	ABEMA (https://abema.tv); Asahi Shimbun (https://www.asahi.com); Fuji TV (https://www.fnn.jp); LINE NEWS (https://news.line.me); Mainichi Shimbun (https://mainichi.jp); NHK News (https://www3.nhk.or.jp/news/); Nikkei (https://www.nikkei.com); Nikkei Business (https://business.nikkei.com); Nippon TV (https://news.ntv.co.jp); Sankei Shimbun (https://www.sankei.com); TBS News (https://newsdig.tbs.co.jp); TV Asahi (https://news.tv-asahi.co.jp); TV Tokyo (https://www.tv-tokyo.co.jp/news/); Toyo Keizai (https://toyokeizai.net); Weekly Bunshun (https://bunshun.jp); Yahoo! Japan News (https://news.yahoo.co.jp); Yomiuri Shimbun (https://www.yomiuri.co.jp)
India	All India Radio (https://www.newsonair.gov.in); CNBC-TV18 (https://www.cnbctv18.com); DD India (https://www.ddindia.co.in); DD News (https://ddnews.gov.in); Dainik Bhaskar (https://www.bhaskar.com); Doordarshan (https://prasarbharati.gov.in); Hindustan Times (https://www.hindustantimes.com); India Today (https://www.indiatoday.in); India.com (https://www.india.com); Malayala Manorama (https://www.manoramaonline.com); NDTV (https://www.ndtv.com); News18 (https://www.news18.com); NewsClick (https://www.newsclick.in); Rediff (https://www.rediff.com); Republic TV (https://www.republicworld.com); The Economic Times (https://economictimes.indiatimes.com); The Hindu (https://www.thehindu.com); The Indian Express (https://indianexpress.com); The New Indian Express (https://www.newindianexpress.com); The Quint (https://www.thequint.com); The Times of India (https://timesofindia.indiatimes.com); ThePrint (https://theprint.in); Times Now (https://www.timesnownews.com)
South Korea	Channel A (https://ichannela.com/news); Chosun Ilbo (https://www.chosun.com); Daum News (https://news.daum.net); Dong-A Ilbo (https://www.donga.com); Hankook Ilbo (https://www.hankookilbo.com); Hankyoreh (https://www.hani.co.kr); JTBC News (https://news.jtbc.co.kr); JoongAng Ilbo (https://www.joongang.co.kr); KBS News (https://news.kbs.co.kr); Korea Economic Daily (https://www.hankyung.com); Kyunghyang Shinmun (https://www.khan.co.kr); MBC News (https://imnews.imbc.com); MBN (https://www.mbn.co.kr/news); Maeil Business Newspaper (https://www.mk.co.kr); Naver News (https://news.naver.com); SBS News (https://news.sbs.co.kr); TV Chosun (https://news.tvchosun.com); YTN (https://www.ytn.co.kr); Yonhap News (https://www.yna.co.kr); Yonhap News TV (https://www.yonhapnewstv.co.kr)

Notes: This table lists the news outlets included in the descriptive analysis together with the corresponding URLs used to identify articles in the CC-News archive. Outlets are drawn from the Reuters Digital News Report and harmonized across survey waves to produce a unique set of outlet names. The outlet–URL pairs define the sampling frame used throughout the descriptive analysis.

Table A.12: News outlets included: South Africa, Nigeria and Kenya

Country	News outlets (URL)
South Africa	BusinessTech (https://businesstech.co.za); City Press (https://www.news24.com/citypress); Daily Maverick (https://www.dailymaverick.co.za); Daily Sun (https://www.snl24.com/dailysun); Eyewitness News (https://www.ewn.co.za); IOL (https://www.iol.co.za); Isolezwe (https://www.iol.co.za/isolezwe); Mail & Guardian (https://mg.co.za); Metro FM (https://www.metrofm.co.za); Netwerk24 (https://www.netwerk24.com); News24 (https://www.news24.com); Newzroom Afrika (https://www.newzroomafrika.tv); SABC News (https://www.sabcnews.com); SAFM (https://www.safm.co.za); Sowetan (https://www.sowetanlive.co.za); Sunday Times (https://www.timeslive.co.za/sunday-times); The Citizen (https://www.citizen.co.za); The South African (https://www.thesouthafrican.com); The Star (https://www.iol.co.za/the-star); TimesLIVE (https://www.timeslive.co.za); Ukhozi FM (https://www.ukhozifm.co.za); eNCA (https://www.enca.com)
Nigeria	AIT (https://ait.live); Arise News (https://www.arise.tv); Channels Television (https://www.channelstv.com); Cool FM (https://coolfm.ng); Daily Post (https://dailypost.ng); Daily Trust (https://dailytrust.com); Information Nigeria (https://www.informationng.com); Legit.ng (https://www.legit.ng); NET (https://thenet.ng); NTA (https://nta.ng); News Agency of Nigeria (https://nannews.ng); News24 Nigeria (https://www.news24.com/news24/africa/nigeria); Nigerian Tribune (https://tribuneonlineng.com); Online Nigeria (https://www.onlinenigeria.com); Premium Times (https://www.premiumtimesng.com); Pulse Nigeria (https://www.pulse.ng); Radio Nigeria (https://radionigeria.gov.ng); Raypower FM (https://raypower.fm); Sahara Reporters (https://saharareporters.com); Silverbird TV (https://silverbirdtv.com); THISDAY (https://www.thisdaylive.com); TVC News (https://www.tvcnews.tv); The Guardian Nigeria (https://guardian.ng); The Nation (https://thenationonlineng.net); The Punch (https://punchng.com); The Sun Nigeria (https://sunnewsonline.com); Vanguard (https://www.vanguardngr.com); Wazobia FM (https://wazobiafm.com)
Kenya	Business Daily (https://www.businessdailyafrica.com); Citizen Digital (https://www.citizen.digital); Citizen Radio (https://www.citizen.digital/radio); Citizen TV (https://www.citizen.digital); Citizen Weekly (https://www.citizen.digital); Classic FM (https://classic105.com); Daily Nation (https://nation.africa); K24 (https://k24.digital); KBC (https://www.kbc.co.ke); KTN News (https://www.standardmedia.co.ke/ktnnews); Kenyans.co.ke (https://www.kenyans.co.ke); NTV Kenya (https://ntvkenya.co.ke); Nairobi News (https://nairobinews.nation.africa); People Daily (https://peopledaily.digital); Radio Jambo (https://radiojambo.co.ke); Radio Maisha (https://www.standardmedia.co.ke/radiomaisha); TV47 (https://www.tv47.digital); Taifa Leo (https://nation.africa/kenya/taifa-leo); The Nairobian (https://www.standardmedia.co.ke/thenairobian); The Standard (https://www.standardmedia.co.ke); The Star (https://www.the-star.co.ke); Tuko (https://www.tuko.co.ke)

Notes: This table lists the news outlets included in the descriptive analysis together with the corresponding URLs used to identify articles in the CC-News archive. Outlets are drawn from the Reuters Digital News Report and harmonized across survey waves to produce a unique set of outlet names. The outlet–URL pairs define the sampling frame used throughout the descriptive analysis.

Table A.13: Descriptive statistics for CC-News HTML population: United Kingdom and Australia

Country	Statistic	Total	2022	2023	2024	2025	2026
United Kingdom	Outlets with coverage	13	13	12	9	4	1
United Kingdom	CC-News HTMLs	10,712,430	4,112,650	3,716,490	2,455,176	380,172	47,942
United Kingdom	Mean HTMLs per outlet	824,033	316,358	309,708	272,797	95,043	47,942
United Kingdom	Median HTMLs per outlet	403,133	163,519	130,595	94,014	34,644	47,942
United Kingdom	Min HTMLs per outlet	4	2	10,854	2,744	2	47,942
United Kingdom	Max HTMLs per outlet	4,989,815	1,719,304	1,751,729	1,518,782	310,882	47,942
Australia	Outlets with coverage	13	11	12	9	2	2
Australia	CC-News HTMLs	2,736,316	1,017,326	1,151,844	523,206	31,266	12,674
Australia	Mean HTMLs per outlet	210,486	92,484	95,987	58,134	15,633	6,337
Australia	Median HTMLs per outlet	118,727	60,702	48,622	31,050	15,633	6,337
Australia	Min HTMLs per outlet	14	8	2	4	2	2
Australia	Max HTMLs per outlet	811,045	288,661	322,242	200,142	31,264	12,672

Notes: This table reports descriptive statistics for the population of CC-News HTMLs matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched HTML in a given country-year. CC-News HTMLs reports the total number of matched HTML records. Mean, median, minimum, and maximum values are computed across outlet-level HTML counts within each country-year. The Total column pools all matched HTMLs from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.14: Descriptive statistics for CC-News HTML population: Germany, France and Italy

Country	Statistic	Total	2022	2023	2024	2025	2026
Germany	Outlets with coverage	15	15	15	7	6	4
Germany	CC-News HTMLs	4,406,571	1,957,499	1,863,520	354,360	195,784	35,408
Germany	Mean HTMLs per outlet	293,771	130,500	124,235	50,623	32,631	8,852
Germany	Median HTMLs per outlet	179,714	83,262	82,922	36,208	32,595	1,894
Germany	Min HTMLs per outlet	16,042	214	2	1,146	130	4
Germany	Max HTMLs per outlet	942,486	494,546	447,940	107,854	60,652	31,616
France	Outlets with coverage	19	19	19	10	7	3
France	CC-News HTMLs	2,811,588	1,309,040	961,984	364,506	143,786	32,272
France	Mean HTMLs per outlet	147,978	68,897	50,631	36,451	20,541	10,757
France	Median HTMLs per outlet	116,406	54,046	46,728	27,296	1,340	14,006
France	Min HTMLs per outlet	6	4	2	4	2	4,194
France	Max HTMLs per outlet	383,959	202,536	126,295	101,072	92,302	14,072
Italy	Outlets with coverage	13	11	13	10	7	4
Italy	CC-News HTMLs	4,307,819	1,387,926	1,376,225	886,606	540,962	116,100
Italy	Mean HTMLs per outlet	331,371	126,175	105,863	88,661	77,280	29,025
Italy	Median HTMLs per outlet	273,300	124,280	71,116	35,101	56,462	21,003
Italy	Min HTMLs per outlet	2	8,872	2	14	22	13,746
Italy	Max HTMLs per outlet	1,349,173	428,340	397,291	410,138	169,500	60,348

Notes: This table reports descriptive statistics for the population of CC-News HTMLs matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched HTML in a given country-year. CC-News HTMLs reports the total number of matched HTML records. Mean, median, minimum, and maximum values are computed across outlet-level HTML counts within each country-year. The Total column pools all matched HTMLs from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.15: Descriptive statistics for CC-News HTML population: United States, Canada and Mexico

Country	Statistic	Total	2022	2023	2024	2025	2026
United States	Outlets with coverage	17	17	16	11	5	6
United States	CC-News HTMLs	5,811,547	1,937,035	1,738,506	1,205,062	637,280	293,664
United States	Mean HTMLs per outlet	341,856	113,943	108,657	109,551	127,456	48,944
United States	Median HTMLs per outlet	157,448	65,804	64,103	30,272	72,190	13,361
United States	Min HTMLs per outlet	10	2	294	134	2	8
United States	Max HTMLs per outlet	2,420,838	448,307	702,405	570,650	474,128	225,348
Canada	Outlets with coverage	15	15	14	7	5	2
Canada	CC-News HTMLs	2,829,503	1,208,533	975,038	382,570	234,792	28,570
Canada	Mean HTMLs per outlet	188,634	80,569	69,646	54,653	46,958	14,285
Canada	Median HTMLs per outlet	183,740	82,218	75,752	45,254	41,632	14,285
Canada	Min HTMLs per outlet	2,446	2,440	6	11,446	2,192	7,694
Canada	Max HTMLs per outlet	380,069	246,405	133,664	124,874	128,882	20,876
Mexico	Outlets with coverage	12	9	11	8	8	7
Mexico	CC-News HTMLs	4,133,540	883,178	1,187,336	1,107,342	614,330	341,354
Mexico	Mean HTMLs per outlet	344,462	98,131	107,940	138,418	76,791	48,765
Mexico	Median HTMLs per outlet	255,445	75,266	74,502	105,016	60,472	46,640
Mexico	Min HTMLs per outlet	2	3,892	2	49,694	46	44
Mexico	Max HTMLs per outlet	837,238	208,622	281,880	332,802	165,750	101,476

Notes: This table reports descriptive statistics for the population of CC-News HTMLs matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched HTML in a given country-year. CC-News HTMLs reports the total number of matched HTML records. Mean, median, minimum, and maximum values are computed across outlet-level HTML counts within each country-year. The Total column pools all matched HTMLs from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.16: Descriptive statistics for CC-News HTML population: Brazil, Argentina and Colombia

Country	Statistic	Total	2022	2023	2024	2025	2026
Brazil	Outlets with coverage	5	3	4	5	5	2
Brazil	CC-News HTMLs	1,234,674	131,426	323,992	525,012	237,578	16,666
Brazil	Mean HTMLs per outlet	246,935	43,809	80,998	105,002	47,516	8,333
Brazil	Median HTMLs per outlet	282,636	44,632	76,737	88,976	30,210	8,333
Brazil	Min HTMLs per outlet	77,342	39,034	33,130	34,924	10,018	40
Brazil	Max HTMLs per outlet	362,054	47,760	137,388	186,646	142,278	16,626
Argentina	Outlets with coverage	17	15	15	15	16	14
Argentina	CC-News HTMLs	11,471,587	2,756,932	2,465,827	2,438,456	2,423,834	1,386,538
Argentina	Mean HTMLs per outlet	674,799	183,795	164,388	162,564	151,490	99,038
Argentina	Median HTMLs per outlet	342,552	100,048	93,742	90,120	83,308	48,634
Argentina	Min HTMLs per outlet	84,422	18,978	42,584	27,986	39,244	26,418
Argentina	Max HTMLs per outlet	3,711,183	898,035	643,538	829,898	804,598	535,114
Colombia	Outlets with coverage	14	12	13	13	12	11
Colombia	CC-News HTMLs	5,073,095	1,504,932	1,218,907	1,133,316	852,608	363,332
Colombia	Mean HTMLs per outlet	362,364	125,411	93,762	87,178	71,051	33,030
Colombia	Median HTMLs per outlet	273,170	67,172	73,186	83,264	70,041	29,064
Colombia	Min HTMLs per outlet	78,294	11,084	22,232	21,576	4	2,672
Colombia	Max HTMLs per outlet	968,514	599,317	242,364	201,966	148,990	81,938

Notes: This table reports descriptive statistics for the population of CC-News HTMLs matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched HTML in a given country-year. CC-News HTMLs reports the total number of matched HTML records. Mean, median, minimum, and maximum values are computed across outlet-level HTML counts within each country-year. The Total column pools all matched HTMLs from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.17: Descriptive statistics for CC-News HTML population: Japan, India and South Korea

Country	Statistic	Total	2022	2023	2024	2025	2026
Japan	Outlets with coverage	5	4	3	3	1	1
Japan	CC-News HTMLs	241,534	113,284	86,232	16,118	15,926	9,974
Japan	Mean HTMLs per outlet	48,307	28,321	28,744	5,373	15,926	9,974
Japan	Median HTMLs per outlet	552	7,525	14,748	96	15,926	9,974
Japan	Min HTMLs per outlet	2	2	410	2	15,926	9,974
Japan	Max HTMLs per outlet	169,306	98,232	71,074	16,020	15,926	9,974
India	Outlets with coverage	17	17	16	11	10	11
India	CC-News HTMLs	13,187,255	4,283,780	3,428,187	2,557,454	1,854,082	1,063,752
India	Mean HTMLs per outlet	775,721	251,987	214,262	232,496	185,408	96,705
India	Median HTMLs per outlet	690,277	235,779	165,119	207,970	185,047	79,444
India	Min HTMLs per outlet	2	2	42,990	42,750	4,438	11,822
India	Max HTMLs per outlet	2,554,500	666,432	660,024	576,066	487,380	336,582
South Korea	Outlets with coverage	7	4	4	7	2	3
South Korea	CC-News HTMLs	9,023,172	2,333,396	1,595,678	3,478,730	1,023,402	591,966
South Korea	Mean HTMLs per outlet	1,289,025	583,349	398,920	496,961	511,701	197,322
South Korea	Median HTMLs per outlet	1,255,225	601,396	323,074	235,622	511,701	130,036
South Korea	Min HTMLs per outlet	8	184,788	112,048	2	268,604	61,152
South Korea	Max HTMLs per outlet	3,698,252	945,816	837,482	2,117,722	754,798	400,778

Notes: This table reports descriptive statistics for the population of CC-News HTMLs matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched HTML in a given country-year. CC-News HTMLs reports the total number of matched HTML records. Mean, median, minimum, and maximum values are computed across outlet-level HTML counts within each country-year. The Total column pools all matched HTMLs from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.18: Descriptive statistics for CC-News HTML population: South Africa, Nigeria and Kenya

Country	Statistic	Total	2022	2023	2024	2025	2026
South Africa	Outlets with coverage	14	11	12	11	8	5
South Africa	CC-News HTMLs	1,480,350	445,134	449,122	277,414	214,850	93,830
South Africa	Mean HTMLs per outlet	105,739	40,467	37,427	25,219	26,856	18,766
South Africa	Median HTMLs per outlet	60,520	35,048	26,727	13,758	23,941	17,422
South Africa	Min HTMLs per outlet	2	2	4	2	6,668	4,988
South Africa	Max HTMLs per outlet	322,306	104,360	131,780	78,144	52,732	30,580
Nigeria	Outlets with coverage	20	18	19	16	13	11
Nigeria	CC-News HTMLs	2,517,754	759,886	731,038	476,054	428,716	122,060
Nigeria	Mean HTMLs per outlet	125,888	42,216	38,476	29,753	32,978	11,096
Nigeria	Median HTMLs per outlet	77,333	42,804	28,172	15,636	25,530	8,692
Nigeria	Min HTMLs per outlet	4	2	2	2	954	2
Nigeria	Max HTMLs per outlet	375,914	111,036	124,892	103,668	74,696	38,358
Kenya	Outlets with coverage	10	9	7	7	9	4
Kenya	CC-News HTMLs	580,894	174,144	146,022	117,400	82,920	60,408
Kenya	Mean HTMLs per outlet	58,089	19,349	20,860	16,771	9,213	15,102
Kenya	Median HTMLs per outlet	18,091	10,682	9,424	7,870	296	15,922
Kenya	Min HTMLs per outlet	12	2	2	4	2	2
Kenya	Max HTMLs per outlet	281,742	76,328	63,454	59,494	56,474	28,562

Notes: This table reports descriptive statistics for the population of CC-News HTMLs matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched HTML in a given country-year. CC-News HTMLs reports the total number of matched HTML records. Mean, median, minimum, and maximum values are computed across outlet-level HTML counts within each country-year. The Total column pools all matched HTMLs from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.19: Descriptive statistics for CC-News random sample: United Kingdom and Australia

Country	Statistic	Total	2022	2023	2024	2025	2026
United Kingdom	Sampled HTMLs	22,200	5,328	5,328	5,328	5,328	888
United Kingdom	Sampled HTMLs with image	15,556	3,691	3,599	3,426	4,096	744
United Kingdom	Sampled HTMLs with image (%)	70.1	69.3	67.5	64.3	76.9	83.8
United Kingdom	Unique images extracted	50,393	11,023	11,078	11,867	15,095	1,330
United Kingdom	Mean images per HTML	2.5	2.2	2.2	2.4	3.4	1.6
United Kingdom	Unique images downloaded	27,689	7,639	7,396	6,475	4,851	1,328
United Kingdom	Unique images download success (%)	54.9	69.3	66.8	54.6	32.1	99.8
Australia	Sampled HTMLs	23,693	5,340	5,340	5,340	5,340	2,333
Australia	Sampled HTMLs with image	17,197	2,881	2,604	4,631	4,976	2,105
Australia	Sampled HTMLs with image (%)	72.6	54.0	48.8	86.7	93.2	90.2
Australia	Unique images extracted	50,018	9,771	9,255	14,411	11,436	5,145
Australia	Mean images per HTML	4.0	3.4	4.2	7.2	2.2	2.2
Australia	Unique images downloaded	43,905	7,747	7,381	12,201	11,434	5,142
Australia	Unique images download success (%)	87.8	79.3	79.8	84.7	100.0	99.9

Notes: This table reports descriptive statistics for the 1% stratified random sample of CC-News HTMLs used for image collection. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled HTMLs. Mean images per HTML reports the mean number of extracted image occurrences per sampled HTML. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.20: Descriptive statistics for CC-News random sample: Germany, France and Italy

Country	Statistic	Total	2022	2023	2024	2025	2026
Germany	Sampled HTMLs	24,033	5,340	5,340	5,340	5,340	2,673
Germany	Sampled HTMLs with image	13,581	2,770	2,876	3,121	3,302	1,512
Germany	Sampled HTMLs with image (%)	56.5	51.9	53.9	58.4	61.8	56.6
Germany	Unique images extracted	42,119	14,566	16,070	4,530	4,763	2,190
Germany	Mean images per HTML	3.2	7.1	5.0	1.1	1.0	0.9
Germany	Unique images downloaded	36,753	12,258	13,482	4,186	4,725	2,102
Germany	Unique images download success (%)	87.3	84.2	83.9	92.4	99.2	96.0
France	Sampled HTMLs	24,031	5,340	5,340	5,340	5,340	2,671
France	Sampled HTMLs with image	12,655	3,055	3,204	2,728	2,462	1,206
France	Sampled HTMLs with image (%)	52.7	57.2	60.0	51.1	46.1	45.2
France	Unique images extracted	31,301	11,108	11,327	4,373	2,797	1,696
France	Mean images per HTML	1.9	2.7	3.1	1.1	1.0	1.0
France	Unique images downloaded	27,019	9,008	9,655	4,344	2,338	1,674
France	Unique images download success (%)	86.3	81.1	85.2	99.3	83.6	98.7
Italy	Sampled HTMLs	24,036	5,340	5,340	5,340	5,340	2,676
Italy	Sampled HTMLs with image	15,616	3,435	3,426	3,325	3,376	2,054
Italy	Sampled HTMLs with image (%)	65.0	64.3	64.2	62.3	63.2	76.8
Italy	Unique images extracted	63,117	15,667	15,105	13,555	11,031	7,759
Italy	Mean images per HTML	5.1	6.0	5.8	4.9	4.0	4.9
Italy	Unique images downloaded	60,688	14,455	14,010	13,438	11,027	7,758
Italy	Unique images download success (%)	96.2	92.3	92.8	99.1	100.0	100.0

Notes: This table reports descriptive statistics for the 1% stratified random sample of CC-News HTMLs used for image collection. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled HTMLs. Mean images per HTML reports the mean number of extracted image occurrences per sampled HTML. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.21: Descriptive statistics for CC-News random sample: United States, Canada and Mexico

Country	Statistic	Total	2022	2023	2024	2025	2026
United States	Sampled HTMLs	24,100	5,328	5,328	5,328	5,328	2,788
United States	Sampled HTMLs with image	6,703	2,273	2,178	1,116	466	670
United States	Sampled HTMLs with image (%)	27.8	42.7	40.9	20.9	8.7	24.0
United States	Unique images extracted	18,013	4,990	6,892	4,324	619	1,188
United States	Mean images per HTML	0.9	1.2	1.6	0.9	0.1	0.4
United States	Unique images downloaded	17,763	4,837	6,802	4,318	619	1,187
United States	Unique images download success (%)	98.6	96.9	98.7	99.9	100.0	99.9
Canada	Sampled HTMLs	24,035	5,340	5,340	5,340	5,340	2,675
Canada	Sampled HTMLs with image	12,777	2,007	2,132	2,870	4,506	1,262
Canada	Sampled HTMLs with image (%)	53.2	37.6	39.9	53.7	84.4	47.2
Canada	Unique images extracted	26,683	3,353	3,688	4,019	14,088	1,535
Canada	Mean images per HTML	1.6	0.6	0.7	1.0	4.7	0.6
Canada	Unique images downloaded	25,799	3,000	3,383	4,016	13,873	1,527
Canada	Unique images download success (%)	96.7	89.5	91.7	99.9	98.5	99.5
Mexico	Sampled HTMLs	24,175	5,340	5,340	5,340	5,340	2,815
Mexico	Sampled HTMLs with image	10,183	1,625	1,462	1,798	3,183	2,115
Mexico	Sampled HTMLs with image (%)	42.1	30.4	27.4	33.7	59.6	75.1
Mexico	Unique images extracted	31,552	2,817	2,064	3,984	14,360	8,327
Mexico	Mean images per HTML	1.9	0.8	0.4	0.8	3.6	5.6
Mexico	Unique images downloaded	29,299	2,645	1,545	3,956	12,829	8,324
Mexico	Unique images download success (%)	92.9	93.9	74.9	99.3	89.3	100.0

Notes: This table reports descriptive statistics for the 1% stratified random sample of CC-News HTMLs used for image collection. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled HTMLs. Mean images per HTML reports the mean number of extracted image occurrences per sampled HTML. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.22: Descriptive statistics for CC-News random sample: Brazil, Argentina and Colombia

Country	Statistic	Total	2022	2023	2024	2025	2026
Brazil	Sampled HTMLs	23,501	5,340	5,340	5,340	5,247	2,234
Brazil	Sampled HTMLs with image	7,899	1,612	1,710	2,087	2,081	409
Brazil	Sampled HTMLs with image (%)	33.6	30.2	32.0	39.1	39.7	18.3
Brazil	Unique images extracted	9,109	2,444	1,541	1,808	2,943	373
Brazil	Mean images per HTML	1.3	0.7	0.8	1.6	2.5	0.2
Brazil	Unique images downloaded	8,652	2,416	1,472	1,693	2,702	369
Brazil	Unique images download success (%)	95.0	98.9	95.5	93.6	91.8	98.9
Argentina	Sampled HTMLs	24,422	5,340	5,340	5,340	5,340	3,062
Argentina	Sampled HTMLs with image	15,175	3,268	3,231	3,400	3,424	1,852
Argentina	Sampled HTMLs with image (%)	62.1	61.2	60.5	63.7	64.1	60.5
Argentina	Unique images extracted	61,356	18,063	14,484	14,170	9,873	4,766
Argentina	Mean images per HTML	8.5	9.2	9.6	10.0	7.2	5.2
Argentina	Unique images downloaded	54,693	16,277	12,760	11,773	9,202	4,681
Argentina	Unique images download success (%)	89.1	90.1	88.1	83.1	93.2	98.2
Colombia	Sampled HTMLs	24,153	5,340	5,340	5,340	5,340	2,793
Colombia	Sampled HTMLs with image	13,581	1,964	2,746	3,472	3,695	1,704
Colombia	Sampled HTMLs with image (%)	56.2	36.8	51.4	65.0	69.2	61.0
Colombia	Unique images extracted	33,464	4,886	5,094	10,358	8,409	4,717
Colombia	Mean images per HTML	2.6	1.3	1.2	3.8	3.6	3.9
Colombia	Unique images downloaded	29,780	2,542	4,117	10,094	8,316	4,711
Colombia	Unique images download success (%)	89.0	52.0	80.8	97.5	98.9	99.9

Notes: This table reports descriptive statistics for the 1% stratified random sample of CC-News HTMLs used for image collection. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled HTMLs. Mean images per HTML reports the mean number of extracted image occurrences per sampled HTML. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.23: Descriptive statistics for CC-News random sample: Japan, India and South Korea

Country	Statistic	Total	2022	2023	2024	2025	2026
Japan	Sampled HTMLs	24,046	5,340	5,340	5,340	5,340	2,686
Japan	Sampled HTMLs with image	4,973	2,666	2,285	22	0	0
Japan	Sampled HTMLs with image (%)	20.7	49.9	42.8	0.4	0.0	0.0
Japan	Unique images extracted	32,036	13,687	17,965	384	0	0
Japan	Mean images per HTML	2.3	4.3	6.1	0.1	0.0	0.0
Japan	Unique images downloaded	29,275	13,117	16,149	9	0	0
Japan	Unique images download success (%)	91.4	95.8	89.9	2.3		
India	Sampled HTMLs	24,316	5,340	5,340	5,340	5,340	2,956
India	Sampled HTMLs with image	9,455	2,204	2,141	2,028	1,782	1,300
India	Sampled HTMLs with image (%)	38.9	41.3	40.1	38.0	33.4	44.0
India	Unique images extracted	35,839	6,537	8,779	10,943	5,273	4,307
India	Mean images per HTML	2.5	2.1	2.6	3.4	2.0	2.0
India	Unique images downloaded	33,070	5,392	8,025	10,599	4,774	4,280
India	Unique images download success (%)	92.3	82.5	91.4	96.9	90.5	99.4
South Korea	Sampled HTMLs	24,040	5,340	5,340	5,340	5,340	2,680
South Korea	Sampled HTMLs with image	10,726	1,926	2,155	3,480	2,373	792
South Korea	Sampled HTMLs with image (%)	44.6	36.1	40.4	65.2	44.4	29.6
South Korea	Unique images extracted	26,078	4,854	6,091	9,681	3,984	1,468
South Korea	Mean images per HTML	5.8	5.2	8.5	11.2	1.0	0.7
South Korea	Unique images downloaded	22,151	4,020	4,431	8,256	3,982	1,462
South Korea	Unique images download success (%)	84.9	82.8	72.7	85.3	99.9	99.6

Notes: This table reports descriptive statistics for the 1% stratified random sample of CC-News HTMLs used for image collection. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled HTMLs. Mean images per HTML reports the mean number of extracted image occurrences per sampled HTML. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.24: Descriptive statistics for CC-News random sample: South Africa, Nigeria and Kenya

Country	Statistic	Total	2022	2023	2024	2025	2026
South Africa	Sampled HTMLs	23,976	5,328	5,328	5,328	5,328	2,664
South Africa	Sampled HTMLs with image	5,054	1,399	958	532	1,013	1,152
South Africa	Sampled HTMLs with image (%)	21.1	26.3	18.0	10.0	19.0	43.2
South Africa	Unique images extracted	9,419	2,138	1,240	1,308	2,150	2,583
South Africa	Mean images per HTML	0.5	0.6	0.4	0.3	0.4	1.0
South Africa	Unique images downloaded	7,341	1,367	823	1,011	1,769	2,371
South Africa	Unique images download success (%)	77.9	63.9	66.4	77.3	82.3	91.8
Nigeria	Sampled HTMLs	24,098	5,340	5,340	5,340	5,340	2,738
Nigeria	Sampled HTMLs with image	10,821	1,925	1,954	2,078	3,028	1,836
Nigeria	Sampled HTMLs with image (%)	44.9	36.0	36.6	38.9	56.7	67.1
Nigeria	Unique images extracted	26,585	3,369	5,486	4,464	8,300	4,966
Nigeria	Mean images per HTML	2.4	1.4	1.9	2.1	3.2	4.0
Nigeria	Unique images downloaded	20,074	1,887	3,375	2,724	7,276	4,812
Nigeria	Unique images download success (%)	75.5	56.0	61.5	61.0	87.7	96.9
Kenya	Sampled HTMLs	23,998	5,340	5,340	5,340	5,222	2,756
Kenya	Sampled HTMLs with image	20,143	4,409	4,742	4,386	4,789	1,817
Kenya	Sampled HTMLs with image (%)	83.9	82.6	88.8	82.1	91.7	65.9
Kenya	Unique images extracted	36,791	7,482	8,673	10,969	7,054	2,613
Kenya	Mean images per HTML	3.0	3.2	4.1	3.9	1.6	1.0
Kenya	Unique images downloaded	26,452	4,387	4,624	7,940	6,890	2,611
Kenya	Unique images download success (%)	71.9	58.6	53.3	72.4	97.7	99.9

Notes: This table reports descriptive statistics for the 1% stratified random sample of CC-News HTMLs used for image collection. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled HTMLs. Mean images per HTML reports the mean number of extracted image occurrences per sampled HTML. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.25: Descriptive statistics for GDELT's VGKG HTML population: United Kingdom and Australia

Country	Statistic	Total	2022	2023	2024
United Kingdom	Outlets with coverage	15	9	13	14
United Kingdom	VGKG Articles	1,149,386	316,483	406,876	426,027
United Kingdom	Mean Articles per outlet	76,626	35,165	31,298	30,430
United Kingdom	Median Articles per outlet	22,114	12,008	10,217	12,364
United Kingdom	Min Articles per outlet	1	19	1	3
United Kingdom	Max Articles per outlet	463,088	144,810	159,322	158,956
Australia	Outlets with coverage	16	13	15	14
Australia	VGKG Articles	378,931	120,730	136,773	121,428
Australia	Mean Articles per outlet	23,683	9,287	9,118	8,673
Australia	Median Articles per outlet	13,447	5,615	6,903	6,606
Australia	Min Articles per outlet	41	41	3	19
Australia	Max Articles per outlet	84,919	32,446	28,763	23,710

Notes: This table reports descriptive statistics for the population of GDELT VGKG articles matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched article in a given country-year. VGKG Articles reports the total number of unique matched article URLs. Mean, median, minimum, and maximum values are computed across outlet-level article counts within each country-year. The Total column pools all matched articles from 2022 through December 2024, not included, and reports the number of unique outlets with coverage over this period.

Table A.26: Descriptive statistics for GDELT's VGKG HTML population: Germany, France and Italy

Country	Statistic	Total	2022	2023	2024
Germany	Outlets with coverage	16	11	16	15
Germany	VGKG Articles	1,230,128	253,122	495,324	481,682
Germany	Mean Articles per outlet	76,883	23,011	30,958	32,112
Germany	Median Articles per outlet	66,560	14,733	27,556	31,065
Germany	Min Articles per outlet	7	11	1	6
Germany	Max Articles per outlet	278,687	86,634	100,359	91,694
France	Outlets with coverage	22	18	18	18
France	VGKG Articles	754,174	243,755	271,511	238,908
France	Mean Articles per outlet	34,281	13,542	15,084	13,273
France	Median Articles per outlet	22,846	8,878	9,537	8,555
France	Min Articles per outlet	1	1	1	1
France	Max Articles per outlet	115,599	49,353	44,779	35,292
Italy	Outlets with coverage	12	8	12	12
Italy	VGKG Articles	313,221	68,410	122,096	122,715
Italy	Mean Articles per outlet	26,102	8,551	10,175	10,226
Italy	Median Articles per outlet	18,955	2,178	8,646	10,436
Italy	Min Articles per outlet	21	28	14	7
Italy	Max Articles per outlet	109,899	40,995	35,499	33,405

Notes: This table reports descriptive statistics for the population of GDELT VGKG articles matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched article in a given country-year. VGKG Articles reports the total number of unique matched article URLs. Mean, median, minimum, and maximum values are computed across outlet-level article counts within each country-year. The Total column pools all matched articles from 2022 through December 2024, not included, and reports the number of unique outlets with coverage over this period.

Table A.27: Descriptive statistics for GDELT's VGKG HTML population: United States, Canada and Mexico

Country	Statistic	Total	2022	2023	2024
United States	Outlets with coverage	17	14	16	15
United States	VGKG Articles	2,169,225	857,113	617,960	694,152
United States	Mean Articles per outlet	127,601	61,222	38,622	46,277
United States	Median Articles per outlet	33,132	17,805	15,971	9,371
United States	Min Articles per outlet	2	2	124	5
United States	Max Articles per outlet	862,716	587,264	275,452	519,368
Canada	Outlets with coverage	15	10	14	14
Canada	VGKG Articles	342,775	52,410	153,103	137,262
Canada	Mean Articles per outlet	22,852	5,241	10,936	9,804
Canada	Median Articles per outlet	19,108	1,424	9,129	7,258
Canada	Min Articles per outlet	7	7	22	1
Canada	Max Articles per outlet	65,793	21,917	26,360	26,244
Mexico	Outlets with coverage	16	14	15	13
Mexico	VGKG Articles	480,852	144,243	181,195	155,414
Mexico	Mean Articles per outlet	30,053	10,303	12,080	11,955
Mexico	Median Articles per outlet	17,084	4,340	3,077	7,799
Mexico	Min Articles per outlet	43	4	3	10
Mexico	Max Articles per outlet	154,732	43,705	61,369	49,658

Notes: This table reports descriptive statistics for the population of GDELT VGKG articles matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched article in a given country-year. VGKG Articles reports the total number of unique matched article URLs. Mean, median, minimum, and maximum values are computed across outlet-level article counts within each country-year. The Total column pools all matched articles from 2022 through December 2024, not included, and reports the number of unique outlets with coverage over this period.

Table A.28: Descriptive statistics for GDELT's VGKG HTML population: Brazil, Argentina and Colombia

Country	Statistic	Total	2022	2023	2024
Brazil	Outlets with coverage	7	6	6	5
Brazil	VGKG Articles	170,459	41,212	62,654	66,593
Brazil	Mean Articles per outlet	24,351	6,869	10,442	13,319
Brazil	Median Articles per outlet	1,783	1,304	1,072	644
Brazil	Min Articles per outlet	20	3	140	17
Brazil	Max Articles per outlet	75,756	30,901	34,429	41,327
Argentina	Outlets with coverage	17	10	15	15
Argentina	VGKG Articles	789,259	131,649	334,794	322,816
Argentina	Mean Articles per outlet	46,427	13,165	22,320	21,521
Argentina	Median Articles per outlet	42,932	11,671	21,021	16,654
Argentina	Min Articles per outlet	2	2	2,942	272
Argentina	Max Articles per outlet	92,964	33,554	45,959	47,966
Colombia	Outlets with coverage	18	8	15	14
Colombia	VGKG Articles	406,418	57,842	169,952	178,624
Colombia	Mean Articles per outlet	22,579	7,230	11,330	12,759
Colombia	Median Articles per outlet	15,120	861	1,941	13,026
Colombia	Min Articles per outlet	1	2	21	1
Colombia	Max Articles per outlet	148,729	44,480	60,674	43,575

Notes: This table reports descriptive statistics for the population of GDELT VGKG articles matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched article in a given country-year. VGKG Articles reports the total number of unique matched article URLs. Mean, median, minimum, and maximum values are computed across outlet-level article counts within each country-year. The Total column pools all matched articles from 2022 through December 2024, not included, and reports the number of unique outlets with coverage over this period.

Table A.29: Descriptive statistics for GDELT's VGKG HTML population: Japan, India and South Korea

Country	Statistic	Total	2022	2023	2024
Japan	Outlets with coverage	9	4	6	7
Japan	VGKG Articles	191,794	39,073	74,335	78,386
Japan	Mean Articles per outlet	21,310	9,768	12,389	11,198
Japan	Median Articles per outlet	1,596	1,608	4,344	1,824
Japan	Min Articles per outlet	1	1	529	1
Japan	Max Articles per outlet	144,317	35,857	52,947	55,513
India	Outlets with coverage	20	10	20	19
India	VGKG Articles	591,232	45,257	278,040	267,935
India	Mean Articles per outlet	29,562	4,526	13,902	14,102
India	Median Articles per outlet	5,472	1,520	410	3,834
India	Min Articles per outlet	3	1	3	2
India	Max Articles per outlet	175,379	25,840	77,765	71,774
South Korea	Outlets with coverage	11	11	8	7
South Korea	VGKG Articles	236,229	70,083	117,483	48,663
South Korea	Mean Articles per outlet	21,475	6,371	14,685	6,952
South Korea	Median Articles per outlet	2,977	2,670	11,582	6,117
South Korea	Min Articles per outlet	4	3	1	24
South Korea	Max Articles per outlet	81,296	25,938	49,418	17,945

Notes: This table reports descriptive statistics for the population of GDELT VGKG articles matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched article in a given country-year. VGKG Articles reports the total number of unique matched article URLs. Mean, median, minimum, and maximum values are computed across outlet-level article counts within each country-year. The Total column pools all matched articles from 2022 through December 2024, not included, and reports the number of unique outlets with coverage over this period.

Table A.30: Descriptive statistics for GDELT's VGKG HTML population: South Africa, Nigeria and Kenya

Country	Statistic	Total	2022	2023	2024
South Africa	Outlets with coverage	14	13	13	13
South Africa	VGKG Articles	282,591	87,634	102,884	92,073
South Africa	Mean Articles per outlet	20,185	6,741	7,914	7,083
South Africa	Median Articles per outlet	16,958	3,437	4,996	6,649
South Africa	Min Articles per outlet	396	274	1	10
South Africa	Max Articles per outlet	52,063	19,286	23,859	19,291
Nigeria	Outlets with coverage	22	14	22	22
Nigeria	VGKG Articles	229,675	72,746	84,473	72,456
Nigeria	Mean Articles per outlet	10,440	5,196	3,840	3,293
Nigeria	Median Articles per outlet	4,826	2,850	1,726	1,929
Nigeria	Min Articles per outlet	147	52	51	19
Nigeria	Max Articles per outlet	56,537	18,415	20,326	17,796
Kenya	Outlets with coverage	9	5	7	8
Kenya	VGKG Articles	87,126	24,197	34,742	28,187
Kenya	Mean Articles per outlet	9,681	4,839	4,963	3,523
Kenya	Median Articles per outlet	2,976	535	1,236	1,628
Kenya	Min Articles per outlet	11	1	3	7
Kenya	Max Articles per outlet	47,723	19,003	16,481	12,239

Notes: This table reports descriptive statistics for the population of GDELT VGKG articles matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched article in a given country-year. VGKG Articles reports the total number of unique matched article URLs. Mean, median, minimum, and maximum values are computed across outlet-level article counts within each country-year. The Total column pools all matched articles from 2022 through December 2024, not included, and reports the number of unique outlets with coverage over this period.

Table A.31: Descriptive statistics for GDELT's VGKG random sample: United Kingdom and Australia

Country	Statistic	Total	2022	2023	2024
United Kingdom	Sampled HTMLs	15,242	5,208	5,208	4,826
United Kingdom	Sampled HTMLs with image	15,137	5,172	5,172	4,793
United Kingdom	Sampled HTMLs with image (%)	99.3	99.3	99.3	99.3
United Kingdom	Unique images extracted	27,271	9,647	8,684	8,940
United Kingdom	Mean images per HTML	1.8	1.9	1.7	1.9
United Kingdom	Unique images downloaded	26,204	9,290	8,367	8,547
United Kingdom	Unique images download success (%)	96.1	96.3	96.3	95.6
Australia	Sampled HTMLs	15,239	5,220	5,220	4,799
Australia	Sampled HTMLs with image	15,135	5,184	5,184	4,767
Australia	Sampled HTMLs with image (%)	99.3	99.3	99.3	99.3
Australia	Unique images extracted	25,188	7,438	9,909	7,841
Australia	Mean images per HTML	1.7	1.4	1.9	1.6
Australia	Unique images downloaded	22,609	6,328	8,957	7,324
Australia	Unique images download success (%)	89.8	85.1	90.4	93.4

Notes: This table reports descriptive statistics for the stratified random sample of GDELT VGKG articles used for image collection. Articles were selected using stratified random sampling by country, news outlet, and publication year. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled articles. Mean images per HTML reports the mean number of extracted image occurrences per sampled article. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.32: Descriptive statistics for GDELT's VGKG random sample: Germany, France and Italy

Country	Statistic	Total	2022	2023	2024
Germany	Sampled HTMLs	15,242	5,220	5,220	4,802
Germany	Sampled HTMLs with image	15,102	5,172	5,172	4,758
Germany	Sampled HTMLs with image (%)	99.1	99.1	99.1	99.1
Germany	Unique images extracted	17,922	5,942	5,741	6,239
Germany	Mean images per HTML	1.2	1.2	1.1	1.3
Germany	Unique images downloaded	9,852	3,003	3,064	3,785
Germany	Unique images download success (%)	55.0	50.5	53.4	60.7
France	Sampled HTMLs	15,233	5,220	5,220	4,793
France	Sampled HTMLs with image	15,093	5,172	5,172	4,749
France	Sampled HTMLs with image (%)	99.1	99.1	99.1	99.1
France	Unique images extracted	17,378	6,316	5,836	5,226
France	Mean images per HTML	1.2	1.2	1.1	1.1
France	Unique images downloaded	14,990	5,685	4,973	4,332
France	Unique images download success (%)	86.3	90.0	85.2	82.9
Italy	Sampled HTMLs	15,212	5,220	5,212	4,780
Italy	Sampled HTMLs with image	15,092	5,172	5,172	4,748
Italy	Sampled HTMLs with image (%)	99.2	99.1	99.2	99.3
Italy	Unique images extracted	19,089	6,588	6,432	6,069
Italy	Mean images per HTML	1.3	1.3	1.3	1.3
Italy	Unique images downloaded	18,462	6,187	6,228	6,047
Italy	Unique images download success (%)	96.7	93.9	96.8	99.6

Notes: This table reports descriptive statistics for the stratified random sample of GDELT VGKG articles used for image collection. Articles were selected using stratified random sampling by country, news outlet, and publication year. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled articles. Mean images per HTML reports the mean number of extracted image occurrences per sampled article. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.33: Descriptive statistics for GDELT's VGKG random sample: United States, Canada and Mexico

Country	Statistic	Total	2022	2023	2024
United States	Sampled HTMLs	15,241	5,208	5,208	4,825
United States	Sampled HTMLs with image	15,136	5,172	5,172	4,792
United States	Sampled HTMLs with image (%)	99.3	99.3	99.3	99.3
United States	Unique images extracted	26,171	8,537	9,094	8,540
United States	Mean images per HTML	1.8	1.8	1.8	1.8
United States	Unique images downloaded	23,634	7,250	8,168	8,216
United States	Unique images download success (%)	90.3	84.9	89.8	96.2
Canada	Sampled HTMLs	15,240	5,220	5,220	4,800
Canada	Sampled HTMLs with image	15,100	5,172	5,172	4,756
Canada	Sampled HTMLs with image (%)	99.1	99.1	99.1	99.1
Canada	Unique images extracted	20,762	6,736	7,041	6,985
Canada	Mean images per HTML	1.4	1.3	1.4	1.5
Canada	Unique images downloaded	16,789	5,404	5,556	5,829
Canada	Unique images download success (%)	80.9	80.2	78.9	83.5
Mexico	Sampled HTMLs	15,221	5,208	5,208	4,805
Mexico	Sampled HTMLs with image	15,116	5,172	5,172	4,772
Mexico	Sampled HTMLs with image (%)	99.3	99.3	99.3	99.3
Mexico	Unique images extracted	20,168	6,495	6,602	7,071
Mexico	Mean images per HTML	1.4	1.3	1.3	1.5
Mexico	Unique images downloaded	13,987	3,223	4,739	6,025
Mexico	Unique images download success (%)	69.4	49.6	71.8	85.2

Notes: This table reports descriptive statistics for the stratified random sample of GDELT VGKG articles used for image collection. Articles were selected using stratified random sampling by country, news outlet, and publication year. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled articles. Mean images per HTML reports the mean number of extracted image occurrences per sampled article. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.34: Descriptive statistics for GDELT's VGKG random sample: Brazil, Argentina and Colombia

Country	Statistic	Total	2022	2023	2024
Brazil	Sampled HTMLs	14,557	5,057	4,710	4,790
Brazil	Sampled HTMLs with image	14,449	5,024	4,679	4,746
Brazil	Sampled HTMLs with image (%)	99.3	99.3	99.3	99.1
Brazil	Unique images extracted	17,437	5,348	5,895	6,194
Brazil	Mean images per HTML	1.2	1.1	1.3	1.3
Brazil	Unique images downloaded	15,443	3,679	5,622	6,142
Brazil	Unique images download success (%)	88.6	68.8	95.4	99.2
Argentina	Sampled HTMLs	15,279	5,220	5,220	4,839
Argentina	Sampled HTMLs with image	15,174	5,184	5,184	4,806
Argentina	Sampled HTMLs with image (%)	99.3	99.3	99.3	99.3
Argentina	Unique images extracted	21,946	7,340	7,191	7,415
Argentina	Mean images per HTML	1.4	1.4	1.4	1.5
Argentina	Unique images downloaded	16,975	5,754	5,504	5,717
Argentina	Unique images download success (%)	77.3	78.4	76.5	77.1
Colombia	Sampled HTMLs	15,246	5,220	5,220	4,806
Colombia	Sampled HTMLs with image	15,106	5,172	5,172	4,762
Colombia	Sampled HTMLs with image (%)	99.1	99.1	99.1	99.1
Colombia	Unique images extracted	23,099	10,409	5,923	6,767
Colombia	Mean images per HTML	1.6	2.1	1.1	1.4
Colombia	Unique images downloaded	20,148	9,529	4,985	5,634
Colombia	Unique images download success (%)	87.2	91.5	84.2	83.3

Notes: This table reports descriptive statistics for the stratified random sample of GDELT VGKG articles used for image collection. Articles were selected using stratified random sampling by country, news outlet, and publication year. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled articles. Mean images per HTML reports the mean number of extracted image occurrences per sampled article. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.35: Descriptive statistics for GDELT's VGKG random sample: Japan, India and South Korea

Country	Statistic	Total	2022	2023	2024
Japan	Sampled HTMLs	15,199	5,208	5,208	4,783
Japan	Sampled HTMLs with image	15,094	5,172	5,172	4,750
Japan	Sampled HTMLs with image (%)	99.3	99.3	99.3	99.3
Japan	Unique images extracted	26,063	11,005	8,200	6,858
Japan	Mean images per HTML	1.8	2.2	1.7	1.5
Japan	Unique images downloaded	23,103	10,371	6,804	5,928
Japan	Unique images download success (%)	88.6	94.2	83.0	86.4
India	Sampled HTMLs	15,230	5,220	5,220	4,790
India	Sampled HTMLs with image	15,090	5,172	5,172	4,746
India	Sampled HTMLs with image (%)	99.1	99.1	99.1	99.1
India	Unique images extracted	22,330	8,926	7,031	6,373
India	Mean images per HTML	1.5	1.7	1.4	1.3
India	Unique images downloaded	14,261	4,395	4,810	5,056
India	Unique images download success (%)	63.9	49.2	68.4	79.3
South Korea	Sampled HTMLs	15,190	5,208	5,208	4,774
South Korea	Sampled HTMLs with image	15,087	5,172	5,172	4,743
South Korea	Sampled HTMLs with image (%)	99.3	99.3	99.3	99.4
South Korea	Unique images extracted	21,188	7,291	7,421	6,476
South Korea	Mean images per HTML	1.6	1.7	1.6	1.4
South Korea	Unique images downloaded	16,338	5,407	5,588	5,343
South Korea	Unique images download success (%)	77.1	74.2	75.3	82.5

Notes: This table reports descriptive statistics for the stratified random sample of GDELT VGKG articles used for image collection. Articles were selected using stratified random sampling by country, news outlet, and publication year. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled articles. Mean images per HTML reports the mean number of extracted image occurrences per sampled article. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.36: Descriptive statistics for GDELT's VGKG random sample: South Africa, Nigeria and Kenya

Country	Statistic	Total	2022	2023	2024
South Africa	Sampled HTMLs	15,193	5,208	5,208	4,777
South Africa	Sampled HTMLs with image	15,088	5,172	5,172	4,744
South Africa	Sampled HTMLs with image (%)	99.3	99.3	99.3	99.3
South Africa	Unique images extracted	18,921	6,814	6,290	5,817
South Africa	Mean images per HTML	1.3	1.4	1.2	1.2
South Africa	Unique images downloaded	13,859	4,210	4,837	4,812
South Africa	Unique images download success (%)	73.2	61.8	76.9	82.7
Nigeria	Sampled HTMLs	15,190	5,208	5,208	4,774
Nigeria	Sampled HTMLs with image	15,085	5,172	5,172	4,741
Nigeria	Sampled HTMLs with image (%)	99.3	99.3	99.3	99.3
Nigeria	Unique images extracted	19,772	6,897	6,837	6,038
Nigeria	Mean images per HTML	1.4	1.4	1.4	1.4
Nigeria	Unique images downloaded	12,836	4,638	4,193	4,005
Nigeria	Unique images download success (%)	64.9	67.2	61.3	66.3
Kenya	Sampled HTMLs	15,127	5,208	5,208	4,711
Kenya	Sampled HTMLs with image	15,025	5,172	5,172	4,681
Kenya	Sampled HTMLs with image (%)	99.3	99.3	99.3	99.4
Kenya	Unique images extracted	19,542	6,807	6,719	6,016
Kenya	Mean images per HTML	1.4	1.4	1.3	1.3
Kenya	Unique images downloaded	14,418	6,258	4,258	3,902
Kenya	Unique images download success (%)	73.8	91.9	63.4	64.9

Notes: This table reports descriptive statistics for the stratified random sample of GDELT VGKG articles used for image collection. Articles were selected using stratified random sampling by country, news outlet, and publication year. Sampled HTMLs refers to the number of sampled news articles. Sampled HTMLs with image reports the number of sampled articles for which at least one non-empty image URL was extracted. Sampled HTMLs with image (%) reports this number as a percentage of all sampled HTMLs. Unique images extracted reports the number of distinct non-empty image URLs identified in the sampled articles. Mean images per HTML reports the mean number of extracted image occurrences per sampled article. Unique images downloaded reports the number of distinct extracted image URLs that were successfully retrieved. Unique images download success (%) reports the percentage of unique extracted image URLs that were successfully downloaded.

Table A.37: Descriptive statistics for Wayback Machine's HTML population: United Kingdom and Australia

Country	Statistic	Total	2022	2023	2024	2025	2026
United Kingdom	Outlets with coverage	13	10	11	13	13	13
United Kingdom	Wayback snapshots	1,320,807	165,400	253,981	400,467	409,494	91,465
United Kingdom	Mean snapshots per outlet	101,601	16,540	23,089	30,805	31,500	7,036
United Kingdom	Median snapshots per outlet	126,522	15,156	24,243	37,556	32,520	5,118
United Kingdom	Min snapshots per outlet	55	333	929	3	16	36
United Kingdom	Max snapshots per outlet	170,734	37,196	45,059	50,606	83,160	27,027
Australia	Outlets with coverage	19	19	18	18	18	17
Australia	Wayback snapshots	1,070,085	145,163	246,021	306,659	319,241	53,001
Australia	Mean snapshots per outlet	56,320	7,640	13,668	17,037	17,736	3,118
Australia	Median snapshots per outlet	44,449	3,366	9,606	2,360	14,798	3,500
Australia	Min snapshots per outlet	71	47	17	28	13	8
Australia	Max snapshots per outlet	152,316	21,060	32,291	46,618	43,717	10,209

Notes: This table reports descriptive statistics for the population of Wayback Machine snapshots matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched snapshot in a given country-year. Wayback snapshots reports the total number of unique archived snapshots. A snapshot is identified by its archived timestamp and original URL. Mean, median, minimum, and maximum values are computed across outlet-level snapshot counts within each country-year. The Total column pools all matched snapshots from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.38: Descriptive statistics for Wayback Machine's HTML population: Germany, France and Italy

Country	Statistic	Total	2022	2023	2024	2025	2026
Germany	Outlets with coverage	16	15	15	16	16	13
Germany	Wayback snapshots	1,522,612	213,265	277,590	480,108	452,871	98,778
Germany	Mean snapshots per outlet	95,163	14,218	18,506	30,007	28,304	7,598
Germany	Median snapshots per outlet	111,308	10,688	20,801	36,444	30,938	7,547
Germany	Min snapshots per outlet	4	18	125	2	2	1
Germany	Max snapshots per outlet	330,777	47,079	51,092	115,543	93,847	23,216
France	Outlets with coverage	24	24	24	23	23	23
France	Wayback snapshots	2,095,499	280,076	398,174	691,247	609,953	116,049
France	Mean snapshots per outlet	87,312	11,670	16,591	30,054	26,520	5,046
France	Median snapshots per outlet	105,262	13,862	19,608	33,221	27,737	4,026
France	Min snapshots per outlet	5	3	2	106	197	1
France	Max snapshots per outlet	185,194	31,049	42,575	89,078	49,106	13,677
Italy	Outlets with coverage	16	16	16	16	16	16
Italy	Wayback snapshots	1,068,998	142,349	215,243	327,029	319,287	65,090
Italy	Mean snapshots per outlet	66,812	8,897	13,453	20,439	19,955	4,068
Italy	Median snapshots per outlet	68,304	10,399	9,464	17,708	20,390	4,030
Italy	Min snapshots per outlet	91	21	25	19	14	4
Italy	Max snapshots per outlet	133,975	16,663	29,268	42,773	42,552	7,869

Notes: This table reports descriptive statistics for the population of Wayback Machine snapshots matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched snapshot in a given country-year. Wayback snapshots reports the total number of unique archived snapshots. A snapshot is identified by its archived timestamp and original URL. Mean, median, minimum, and maximum values are computed across outlet-level snapshot counts within each country-year. The Total column pools all matched snapshots from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.39: Descriptive statistics for Wayback Machine's HTML population: United States, Canada and Mexico

Country	Statistic	Total	2022	2023	2024	2025	2026
United States	Outlets with coverage	18	17	17	18	18	17
United States	Wayback snapshots	2,610,969	470,846	542,678	667,398	710,706	219,341
United States	Mean snapshots per outlet	145,054	27,697	31,922	37,078	39,484	12,902
United States	Median snapshots per outlet	141,540	23,274	31,243	38,698	31,178	6,967
United States	Min snapshots per outlet	10	2,057	1,686	8	2	3,236
United States	Max snapshots per outlet	359,930	66,812	63,992	77,191	157,037	75,695
Canada	Outlets with coverage	16	16	16	16	16	15
Canada	Wayback snapshots	816,313	106,079	136,530	226,649	284,850	62,205
Canada	Mean snapshots per outlet	51,020	6,630	8,533	14,166	17,803	4,147
Canada	Median snapshots per outlet	15,546	2,680	1,420	2,347	6,430	3,739
Canada	Min snapshots per outlet	962	353	48	34	13	22
Canada	Max snapshots per outlet	218,642	24,781	37,885	63,473	80,580	11,923
Mexico	Outlets with coverage	18	18	18	18	17	17
Mexico	Wayback snapshots	1,078,148	84,089	132,033	280,713	417,808	163,505
Mexico	Mean snapshots per outlet	59,897	4,672	7,335	15,595	24,577	9,618
Mexico	Median snapshots per outlet	26,158	2,272	1,038	3,834	12,689	3,742
Mexico	Min snapshots per outlet	307	62	41	67	42	17
Mexico	Max snapshots per outlet	234,987	13,254	32,320	44,778	114,294	94,050

Notes: This table reports descriptive statistics for the population of Wayback Machine snapshots matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched snapshot in a given country-year. Wayback snapshots reports the total number of unique archived snapshots. A snapshot is identified by its archived timestamp and original URL. Mean, median, minimum, and maximum values are computed across outlet-level snapshot counts within each country-year. The Total column pools all matched snapshots from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.40: Descriptive statistics for Wayback Machine's HTML population: Brazil, Argentina and Colombia

Country	Statistic	Total	2022	2023	2024	2025	2026
Brazil	Outlets with coverage	19	19	19	18	17	16
Brazil	Wayback snapshots	661,469	117,512	90,617	111,781	246,363	95,196
Brazil	Mean snapshots per outlet	34,814	6,185	4,769	6,210	14,492	5,950
Brazil	Median snapshots per outlet	28,934	3,188	1,331	1,147	10,075	4,508
Brazil	Min snapshots per outlet	627	267	169	62	3	1
Brazil	Max snapshots per outlet	145,356	21,434	28,825	42,010	51,394	14,488
Argentina	Outlets with coverage	22	22	22	21	22	20
Argentina	Wayback snapshots	1,427,748	145,585	246,553	493,879	460,437	81,294
Argentina	Mean snapshots per outlet	64,898	6,618	11,207	23,518	20,929	4,065
Argentina	Median snapshots per outlet	67,120	3,234	5,974	28,325	24,036	3,516
Argentina	Min snapshots per outlet	9	1	1	64	7	3,029
Argentina	Max snapshots per outlet	136,117	17,623	34,527	44,744	40,193	7,826
Colombia	Outlets with coverage	23	21	22	23	23	23
Colombia	Wayback snapshots	657,146	86,089	89,296	147,865	240,404	93,492
Colombia	Mean snapshots per outlet	28,572	4,099	4,059	6,429	10,452	4,065
Colombia	Median snapshots per outlet	16,548	1,667	718	1,194	6,828	3,478
Colombia	Min snapshots per outlet	229	44	16	29	16	4
Colombia	Max snapshots per outlet	114,827	20,943	25,161	36,804	34,232	10,176

Notes: This table reports descriptive statistics for the population of Wayback Machine snapshots matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched snapshot in a given country-year. Wayback snapshots reports the total number of unique archived snapshots. A snapshot is identified by its archived timestamp and original URL. Mean, median, minimum, and maximum values are computed across outlet-level snapshot counts within each country-year. The Total column pools all matched snapshots from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.41: Descriptive statistics for Wayback Machine's HTML population: Japan, India and South Korea

Country	Statistic	Total	2022	2023	2024	2025	2026
Japan	Outlets with coverage	15	15	15	15	15	15
Japan	Wayback snapshots	1,386,192	327,035	147,750	306,469	521,920	83,018
Japan	Mean snapshots per outlet	92,413	21,802	9,850	20,431	34,795	5,535
Japan	Median snapshots per outlet	77,041	4,966	4,343	18,060	24,211	4,039
Japan	Min snapshots per outlet	10,059	286	237	205	5,898	4
Japan	Max snapshots per outlet	329,208	231,811	30,405	43,833	214,655	25,737
India	Outlets with coverage	23	23	23	23	23	23
India	Wayback snapshots	1,854,278	252,983	335,813	559,281	605,888	100,313
India	Mean snapshots per outlet	80,621	10,999	14,601	24,317	26,343	4,361
India	Median snapshots per outlet	77,449	10,372	5,758	29,817	26,171	4,157
India	Min snapshots per outlet	334	5	2	67	33	7
India	Max snapshots per outlet	226,927	32,910	53,446	70,169	125,838	11,191
South Korea	Outlets with coverage	19	19	19	19	19	19
South Korea	Wayback snapshots	1,076,368	98,353	195,383	331,084	383,022	68,526
South Korea	Mean snapshots per outlet	56,651	5,176	10,283	17,425	20,159	3,607
South Korea	Median snapshots per outlet	32,771	2,740	1,638	3,945	17,891	3,416
South Korea	Min snapshots per outlet	185	32	30	53	48	22
South Korea	Max snapshots per outlet	129,680	19,612	29,249	43,324	42,341	9,217

Notes: This table reports descriptive statistics for the population of Wayback Machine snapshots matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched snapshot in a given country-year. Wayback snapshots reports the total number of unique archived snapshots. A snapshot is identified by its archived timestamp and original URL. Mean, median, minimum, and maximum values are computed across outlet-level snapshot counts within each country-year. The Total column pools all matched snapshots from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.42: Descriptive statistics for Wayback Machine's HTML population: South Africa, Nigeria and Kenya

Country	Statistic	Total	2022	2023	2024	2025	2026
South Africa	Outlets with coverage	17	14	13	15	14	12
South Africa	Wayback snapshots	1,230,830	177,795	299,429	394,876	312,379	46,351
South Africa	Mean snapshots per outlet	72,402	12,700	23,033	26,325	22,313	3,863
South Africa	Median snapshots per outlet	114,558	16,283	29,313	42,424	25,916	3,632
South Africa	Min snapshots per outlet	3	11	26	3	18	9
South Africa	Max snapshots per outlet	145,156	20,325	30,799	44,269	44,080	7,684
Nigeria	Outlets with coverage	27	27	27	27	25	24
Nigeria	Wayback snapshots	1,775,566	249,844	371,657	575,433	503,142	75,490
Nigeria	Mean snapshots per outlet	65,762	9,253	13,765	21,312	20,126	3,145
Nigeria	Median snapshots per outlet	34,638	11,083	7,069	4,291	16,213	3,385
Nigeria	Min snapshots per outlet	105	27	19	21	26	9
Nigeria	Max snapshots per outlet	168,004	22,397	30,358	68,680	65,358	7,014
Kenya	Outlets with coverage	15	13	13	15	14	12
Kenya	Wayback snapshots	627,405	74,886	118,253	214,498	189,761	30,007
Kenya	Mean snapshots per outlet	41,827	5,760	9,096	14,300	13,554	2,501
Kenya	Median snapshots per outlet	4,777	2,153	182	530	1,093	2,666
Kenya	Min snapshots per outlet	92	24	2	2	2	8
Kenya	Max snapshots per outlet	124,123	17,077	29,862	42,095	45,710	10,184

Notes: This table reports descriptive statistics for the population of Wayback Machine snapshots matched to the news outlets included in the study. Outlets with coverage refers to the number of outlet domains with at least one matched snapshot in a given country-year. Wayback snapshots reports the total number of unique archived snapshots. A snapshot is identified by its archived timestamp and original URL. Mean, median, minimum, and maximum values are computed across outlet-level snapshot counts within each country-year. The Total column pools all matched snapshots from 2022 through 2026 and reports the number of unique outlets with coverage over this period.

Table A.43: Descriptive statistics for Wayback Machine's random sample: United Kingdom and Australia

Country	Statistic	Total	2022	2023	2024	2025	2026
United Kingdom	Sampled HTMLs	23,536	5,124	5,124	5,124	5,124	3,040
United Kingdom	Sampled HTMLs with image	3,347	1,044	978	817	335	173
United Kingdom	Sampled HTMLs with image (%)	14.2	20.4	19.1	15.9	6.5	5.7
United Kingdom	Unique images extracted	200,710	71,314	64,215	43,853	13,987	7,341
United Kingdom	Mean images per HTML	10.1	15.2	14.4	12.5	2.8	2.4
United Kingdom	Unique images downloaded	50,597	18,882	16,667	10,650	2,660	1,738
United Kingdom	Unique images download success (%)	25.2	26.5	26.0	24.3	19.0	23.7
Australia	Sampled HTMLs	23,573	5,136	5,136	5,136	5,136	3,029
Australia	Sampled HTMLs with image	9,993	1,953	2,275	1,978	1,985	1,802
Australia	Sampled HTMLs with image (%)	42.4	38.0	44.3	38.5	38.6	59.5
Australia	Unique images extracted	673,004	107,240	144,679	135,736	148,951	136,398
Australia	Mean images per HTML	29.9	22.1	29.6	27.7	30.3	46.5
Australia	Unique images downloaded	159,787	23,818	32,668	33,030	36,327	33,944
Australia	Unique images download success (%)	23.7	22.2	22.6	24.3	24.4	24.9

Notes: This table reports descriptive statistics for the stratified random sample of Wayback Machine snapshots of newspaper front pages. Sampled HTMLs refers to the number of selected archived snapshots. Sampled HTMLs with image reports the number of sampled snapshots from which at least one non-empty image URL was extracted from a figure element. Sampled HTMLs with image (%) expresses this count as a percentage of all sampled snapshots. Unique images extracted reports the number of distinct combinations of archival timestamp and non-empty image URL. Mean images per HTML reports the mean number of extracted image occurrences per sampled snapshot. Unique images downloaded reports the number of distinct timestamp-image URL combinations successfully retrieved, either from the Wayback Machine or, when unavailable, from the live URL. Unique images download success (%) reports the corresponding percentage of extracted timestamp-image URL combinations.

Table A.44: Descriptive statistics for Wayback Machine's random sample: Germany, France and Italy

Country	Statistic	Total	2022	2023	2024	2025	2026
Germany	Sampled HTMLs	23,637	5,136	5,136	5,136	5,136	3,093
Germany	Sampled HTMLs with image	11,640	2,622	2,644	2,238	2,620	1,516
Germany	Sampled HTMLs with image (%)	49.2	51.1	51.5	43.6	51.0	49.0
Germany	Unique images extracted	1,244,853	253,694	252,409	233,962	315,996	188,792
Germany	Mean images per HTML	57.6	53.0	52.7	47.9	68.8	70.6
Germany	Unique images downloaded	311,147	64,436	63,666	58,704	75,651	48,690
Germany	Unique images download success (%)	25.0	25.4	25.2	25.1	23.9	25.8
France	Sampled HTMLs	23,566	5,136	5,136	5,136	5,136	3,022
France	Sampled HTMLs with image	9,686	2,146	2,240	2,392	1,859	1,049
France	Sampled HTMLs with image (%)	41.1	41.8	43.6	46.6	36.2	34.7
France	Unique images extracted	439,702	103,715	113,390	101,778	75,679	45,140
France	Mean images per HTML	19.8	21.3	23.6	21.2	15.6	15.5
France	Unique images downloaded	108,729	25,121	28,782	25,133	18,392	11,301
France	Unique images download success (%)	24.7	24.2	25.4	24.7	24.3	25.0
Italy	Sampled HTMLs	23,559	5,136	5,136	5,136	5,136	3,015
Italy	Sampled HTMLs with image	14,357	3,742	3,538	2,983	2,502	1,592
Italy	Sampled HTMLs with image (%)	60.9	72.9	68.9	58.1	48.7	52.8
Italy	Unique images extracted	1,312,981	309,829	306,092	273,356	251,934	171,770
Italy	Mean images per HTML	59.9	65.8	64.4	57.7	51.8	59.5
Italy	Unique images downloaded	367,716	88,141	87,428	76,113	68,590	47,444
Italy	Unique images download success (%)	28.0	28.4	28.6	27.8	27.2	27.6

Notes: This table reports descriptive statistics for the stratified random sample of Wayback Machine snapshots of newspaper front pages. Sampled HTMLs refers to the number of selected archived snapshots. Sampled HTMLs with image reports the number of sampled snapshots from which at least one non-empty image URL was extracted from a figure element. Sampled HTMLs with image (%) expresses this count as a percentage of all sampled snapshots. Unique images extracted reports the number of distinct combinations of archival timestamp and non-empty image URL. Mean images per HTML reports the mean number of extracted image occurrences per sampled snapshot. Unique images downloaded reports the number of distinct timestamp-image URL combinations successfully retrieved, either from the Wayback Machine or, when unavailable, from the live URL. Unique images download success (%) reports the corresponding percentage of extracted timestamp-image URL combinations.

Table A.45: Descriptive statistics for Wayback Machine's random sample: United States, Canada and Mexico

Country	Statistic	Total	2022	2023	2024	2025	2026
United States	Sampled HTMLs	23,576	5,124	5,124	5,124	5,124	3,080
United States	Sampled HTMLs with image	7,817	1,407	1,468	1,898	1,847	1,197
United States	Sampled HTMLs with image (%)	33.2	27.5	28.6	37.0	36.0	38.9
United States	Unique images extracted	111,774	19,607	20,642	21,634	23,368	26,523
United States	Mean images per HTML	5.0	4.2	4.3	4.5	4.8	8.9
United States	Unique images downloaded	31,508	5,528	5,815	6,051	6,575	7,539
United States	Unique images download success (%)	28.2	28.2	28.2	28.0	28.1	28.4
Canada	Sampled HTMLs	23,575	5,136	5,136	5,136	5,136	3,031
Canada	Sampled HTMLs with image	6,034	380	490	1,074	2,366	1,724
Canada	Sampled HTMLs with image (%)	25.6	7.4	9.5	20.9	46.1	56.9
Canada	Unique images extracted	166,974	9,920	10,926	20,664	73,302	52,162
Canada	Mean images per HTML	7.3	1.9	2.1	4.1	15.0	17.8
Canada	Unique images downloaded	46,801	2,904	3,093	5,748	20,442	14,614
Canada	Unique images download success (%)	28.0	29.3	28.3	27.8	27.9	28.0
Mexico	Sampled HTMLs	23,774	5,136	5,136	5,136	5,136	3,230
Mexico	Sampled HTMLs with image	11,750	1,567	2,089	2,614	3,227	2,253
Mexico	Sampled HTMLs with image (%)	49.4	30.5	40.7	50.9	62.8	69.8
Mexico	Unique images extracted	355,359	35,919	48,369	92,589	115,042	63,440
Mexico	Mean images per HTML	17.2	7.0	9.4	18.4	27.2	27.6
Mexico	Unique images downloaded	84,883	8,093	7,540	20,300	30,780	18,170
Mexico	Unique images download success (%)	23.9	22.5	15.6	21.9	26.8	28.6

Notes: This table reports descriptive statistics for the stratified random sample of Wayback Machine snapshots of newspaper front pages. Sampled HTMLs refers to the number of selected archived snapshots. Sampled HTMLs with image reports the number of sampled snapshots from which at least one non-empty image URL was extracted from a figure element. Sampled HTMLs with image (%) expresses this count as a percentage of all sampled snapshots. Unique images extracted reports the number of distinct combinations of archival timestamp and non-empty image URL. Mean images per HTML reports the mean number of extracted image occurrences per sampled snapshot. Unique images downloaded reports the number of distinct timestamp-image URL combinations successfully retrieved, either from the Wayback Machine or, when unavailable, from the live URL. Unique images download success (%) reports the corresponding percentage of extracted timestamp-image URL combinations.

Table A.46: Descriptive statistics for Wayback Machine's random sample: Brazil, Argentina and Colombia

Country	Statistic	Total	2022	2023	2024	2025	2026
Brazil	Sampled HTMLs	23,559	5,136	5,136	5,136	5,136	3,015
Brazil	Sampled HTMLs with image	13,623	2,644	2,806	3,220	2,962	1,991
Brazil	Sampled HTMLs with image (%)	57.8	51.5	54.6	62.7	57.7	66.0
Brazil	Unique images extracted	695,809	101,787	127,925	164,058	170,285	131,754
Brazil	Mean images per HTML	30.1	20.0	25.8	32.4	33.7	44.7
Brazil	Unique images downloaded	176,948	24,983	32,696	39,730	43,346	36,193
Brazil	Unique images download success (%)	25.4	24.5	25.6	24.2	25.5	27.5
Argentina	Sampled HTMLs	23,553	5,136	5,136	5,136	5,136	3,009
Argentina	Sampled HTMLs with image	12,364	3,221	2,531	2,686	2,427	1,499
Argentina	Sampled HTMLs with image (%)	52.5	62.7	49.3	52.3	47.3	49.8
Argentina	Unique images extracted	518,845	137,314	125,263	113,646	94,924	47,698
Argentina	Mean images per HTML	35.2	45.4	36.9	35.3	30.1	23.9
Argentina	Unique images downloaded	128,299	33,064	30,122	27,342	24,410	13,361
Argentina	Unique images download success (%)	24.7	24.1	24.0	24.1	25.7	28.0
Colombia	Sampled HTMLs	23,553	5,136	5,136	5,136	5,136	3,009
Colombia	Sampled HTMLs with image	8,897	1,691	1,941	1,744	2,037	1,484
Colombia	Sampled HTMLs with image (%)	37.8	32.9	37.8	34.0	39.7	49.3
Colombia	Unique images extracted	549,076	69,455	95,212	124,225	147,231	112,953
Colombia	Mean images per HTML	27.3	14.2	21.7	28.6	33.3	46.5
Colombia	Unique images downloaded	152,223	18,097	25,960	34,652	41,313	32,201
Colombia	Unique images download success (%)	27.7	26.1	27.3	27.9	28.1	28.5

Notes: This table reports descriptive statistics for the stratified random sample of Wayback Machine snapshots of newspaper front pages. Sampled HTMLs refers to the number of selected archived snapshots. Sampled HTMLs with image reports the number of sampled snapshots from which at least one non-empty image URL was extracted from a figure element. Sampled HTMLs with image (%) expresses this count as a percentage of all sampled snapshots. Unique images extracted reports the number of distinct combinations of archival timestamp and non-empty image URL. Mean images per HTML reports the mean number of extracted image occurrences per sampled snapshot. Unique images downloaded reports the number of distinct timestamp-image URL combinations successfully retrieved, either from the Wayback Machine or, when unavailable, from the live URL. Unique images download success (%) reports the corresponding percentage of extracted timestamp-image URL combinations.

Table A.47: Descriptive statistics for Wayback Machine's random sample: Japan, India and South Korea

Country	Statistic	Total	2022	2023	2024	2025	2026
Japan	Sampled HTMLs	23,557	5,136	5,136	5,136	5,136	3,013
Japan	Sampled HTMLs with image	7,603	1,545	1,726	1,762	1,641	929
Japan	Sampled HTMLs with image (%)	32.3	30.1	33.6	34.3	32.0	30.8
Japan	Unique images extracted	566,270	94,727	105,816	137,037	131,923	96,767
Japan	Mean images per HTML	31.2	24.4	26.2	34.4	33.5	41.6
Japan	Unique images downloaded	149,293	24,901	28,154	35,281	34,444	26,513
Japan	Unique images download success (%)	26.4	26.3	26.6	25.7	26.1	27.4
India	Sampled HTMLs	23,580	5,136	5,136	5,136	5,136	3,036
India	Sampled HTMLs with image	11,618	2,558	2,674	2,674	2,232	1,480
India	Sampled HTMLs with image (%)	49.3	49.8	52.1	52.1	43.5	48.7
India	Unique images extracted	663,225	124,449	150,156	170,331	149,032	69,257
India	Mean images per HTML	31.0	28.9	30.2	34.6	31.2	30.0
India	Unique images downloaded	185,658	33,837	42,632	47,773	41,878	19,538
India	Unique images download success (%)	28.0	27.2	28.4	28.0	28.1	28.2
South Korea	Sampled HTMLs	23,551	5,136	5,136	5,136	5,136	3,007
South Korea	Sampled HTMLs with image	6,957	1,498	1,324	1,442	1,384	1,309
South Korea	Sampled HTMLs with image (%)	29.5	29.2	25.8	28.1	26.9	43.5
South Korea	Unique images extracted	513,444	97,591	105,069	104,764	115,434	90,586
South Korea	Mean images per HTML	27.8	20.4	22.2	21.9	24.0	66.6
South Korea	Unique images downloaded	137,437	27,032	27,799	27,824	29,794	24,988
South Korea	Unique images download success (%)	26.8	27.7	26.5	26.6	25.8	27.6

Notes: This table reports descriptive statistics for the stratified random sample of Wayback Machine snapshots of newspaper front pages. Sampled HTMLs refers to the number of selected archived snapshots. Sampled HTMLs with image reports the number of sampled snapshots from which at least one non-empty image URL was extracted from a figure element. Sampled HTMLs with image (%) expresses this count as a percentage of all sampled snapshots. Unique images extracted reports the number of distinct combinations of archival timestamp and non-empty image URL. Mean images per HTML reports the mean number of extracted image occurrences per sampled snapshot. Unique images downloaded reports the number of distinct timestamp-image URL combinations successfully retrieved, either from the Wayback Machine or, when unavailable, from the live URL. Unique images download success (%) reports the corresponding percentage of extracted timestamp-image URL combinations.

Table A.48: Descriptive statistics for Wayback Machine's random sample: South Africa, Nigeria and Kenya

Country	Statistic	Total	2022	2023	2024	2025	2026
South Africa	Sampled HTMLs	23,499	5,124	5,124	5,124	5,124	3,003
South Africa	Sampled HTMLs with image	649	110	23	97	93	326
South Africa	Sampled HTMLs with image (%)	2.8	2.1	0.4	1.9	1.8	10.9
South Africa	Unique images extracted	5,603	220	23	746	372	4,242
South Africa	Mean images per HTML	0.2	0.0	0.0	0.2	0.1	1.4
South Africa	Unique images downloaded	1,532	43	10	203	119	1,157
South Africa	Unique images download success (%)	27.3	19.5	43.5	27.2	32.0	27.3
Nigeria	Sampled HTMLs	23,064	5,136	5,136	5,136	5,136	2,520
Nigeria	Sampled HTMLs with image	9,243	1,846	1,790	1,856	2,302	1,449
Nigeria	Sampled HTMLs with image (%)	40.1	35.9	34.9	36.1	44.8	57.5
Nigeria	Unique images extracted	192,794	27,882	43,336	39,596	56,962	25,018
Nigeria	Mean images per HTML	9.1	5.6	9.2	8.6	12.2	10.7
Nigeria	Unique images downloaded	52,113	7,478	11,348	10,551	15,772	6,964
Nigeria	Unique images download success (%)	27.0	26.8	26.2	26.6	27.7	27.8
Kenya	Sampled HTMLs	21,571	5,136	4,905	5,136	5,136	1,258
Kenya	Sampled HTMLs with image	4,375	1,275	877	701	1,194	328
Kenya	Sampled HTMLs with image (%)	20.3	24.8	17.9	13.6	23.2	26.1
Kenya	Unique images extracted	76,978	14,710	4,072	23,008	26,610	8,578
Kenya	Mean images per HTML	4.4	3.6	2.2	5.3	5.6	7.0
Kenya	Unique images downloaded	20,859	3,317	1,157	6,413	7,597	2,375
Kenya	Unique images download success (%)	27.1	22.5	28.4	27.9	28.5	27.7

Notes: This table reports descriptive statistics for the stratified random sample of Wayback Machine snapshots of newspaper front pages. Sampled HTMLs refers to the number of selected archived snapshots. Sampled HTMLs with image reports the number of sampled snapshots from which at least one non-empty image URL was extracted from a figure element. Sampled HTMLs with image (%) expresses this count as a percentage of all sampled snapshots. Unique images extracted reports the number of distinct combinations of archival timestamp and non-empty image URL. Mean images per HTML reports the mean number of extracted image occurrences per sampled snapshot. Unique images downloaded reports the number of distinct timestamp-image URL combinations successfully retrieved, either from the Wayback Machine or, when unavailable, from the live URL. Unique images download success (%) reports the corresponding percentage of extracted timestamp-image URL combinations.

A.7 AI Caption Detection

A.7.1 Prevalence of AI Disclosures in Image Captions

Table A.49 reports the estimated prevalence of AI-related disclosures in image captions in the CC-News sample, discussed in Section 2. The table summarizes the number of figure occurrences, captioned occurrences, and analyzable captions before reporting disclosure rates under each of the four nested disclosure definitions described below.

Table A.49: Estimated prevalence of AI disclosures in image captions

Statistic	CC-News
Sampled articles	406,853
Downloaded image occurrences	956,911
Downloaded image occurrences with captions	230,773
Captions eligible for English-language analysis	215,638
AI disclosures (count)	
Strict AI-image disclosure	164
AI editing disclosure	1
Any AI disclosure	165
Broad AI reference	460
AI disclosures (weighted % of analyzable captions)	
Strict AI-image disclosure	0.09
AI editing disclosure	0.00
Any AI disclosure	0.09
Broad AI reference	0.43
AI disclosures (weighted % of captions deemed AI by 3+ detectors)	
Strict AI-image disclosure	0.21
AI editing disclosure	0.00
Any AI disclosure	0.21
Broad AI reference	0.53

Notes: This table reports aggregate statistics on AI disclosures in captions associated with downloaded image occurrences in the CC-News sample. The sample is restricted to downloaded image occurrences associated with a nonempty figure caption, matching the CC-News sample used in the AI-image prevalence analysis. Disclosure classifications are based on English-language captions and captions translated into English. Strict AI-image disclosures identify captions indicating that an image was AI-generated or naming an AI image-generation model. Any AI disclosure additionally includes AI editing disclosures, while broad AI references also include ambiguous AI references not identified as exclusions. The first set of weighted percentages is calculated among captioned downloaded image occurrences with nonmissing scores from all four detectors whose captions are eligible for English-language disclosure detection. The subsequent set is calculated among these eligible occurrences classified as AI-generated by at least three of the four AI-image detectors. A detector classifies an image as AI-generated when its score is greater than or equal to 0.50. All weighted percentages use the inverse article-sampling probability associated with each downloaded image occurrence.

Table A.50 reports the one-word and two-word phrases that occur most frequently in captions classified as disclosing AI use and in the remaining captions, providing a descriptive comparison of the language used in each group.

Table A.50: Top one-word and two-word phrases by AI and non-AI captions

	Unigrams		Bigrams	
Rank	AI reference	Other	AI reference	Other
1	image	photo	artificial intelligence	getty images
2	ai	credit	representative image	photo credit
3	photo	getty	representational image	abc news
4	artificial	images	photo credit	canadian press
5	intelligence	source	image photo	ap photo
6	representative	new	generated ai	donald trump
7	credit	president	istockphoto com	credit getty
8	generated	afp	image generated	source facebook
9	representational	news	midjourney pulzo	source getty
10	com	picture	pulzo ia	prime minister

Notes: This table reports the ten most frequent unigrams and bigrams in analyzable CC-News image captions. The sample is restricted to downloaded image occurrences associated with a nonempty figure caption and with nonmissing scores from all four AI-image detectors, matching the detector-complete caption sample used in the other caption analyses. AI-reference captions are captions containing a broad AI reference, including strict AI-image disclosures, AI editing disclosures, and ambiguous AI references not identified as exclusions. Other captions are analyzable captions containing no such AI reference. Caption text is based on English-language captions and captions translated into English. Text is normalized to lowercase, punctuation is removed, and standard English stopwords are excluded. Rankings are based on unweighted n-gram frequencies.

A.7.2 Caption Dataset Construction

The detection of AI disclosures begins with the construction of a comprehensive caption dataset from the CC-News sample. For every sampled figure occurrence, we retain the associated figure caption together with auxiliary textual metadata extracted during HTML parsing, including the image alternative text, HTML surrounding the figure element, and image attributes. Throughout the disclosure analysis, figure captions constitute the primary source of textual information, while the remaining fields are retained for diagnostic purposes and future validation.

Before analysis, captions are standardized to Unicode Normalization Form KC (NFKC), transformed to lowercase, and redundant whitespace is removed. Each normalized caption is then assigned a SHA-256 hash, allowing identical captions reused across multiple articles or news outlets to be identified. Because syndicated news images frequently appear with identical captions across publishers, deduplicating captions reduces the computational burden of language

identification and machine translation while ensuring that identical captions receive identical classifications.

Since our sample spans seventeen countries and multiple languages, disclosure detection is performed on a common English representation of every caption. We first identify the language of each unique caption using the `Lingua` language-identification Python library, which assigns both the most likely language and an associated confidence score. Languages representing less than one percent of unique captions are excluded from translation because they account for only a negligible fraction of the corpus.

Captions originally written in English are retained unchanged. Captions written in supported non-English languages are translated into English using Meta's *No Language Left Behind* (`NLLB-200 Distilled 600M`) multilingual sequence-to-sequence translation model. Translation is performed separately for each detected source language. Following translation, captions are normalized using the same preprocessing procedure applied to the original English captions. Captions whose language cannot be reliably identified or translated are excluded from the disclosure analysis.

The resulting dataset contains one normalized English representation for every caption eligible for disclosure detection. Translation metadata are subsequently merged back to every figure occurrence in the original sample, allowing disclosure classifications to be linked directly to the downloaded images and the AI-image detector outputs used throughout the paper.

A.7.3 Rule-Based Detection of AI Disclosures

We identify AI disclosures using a rule-based text-classification procedure designed for image captions. Rather than relying on a single list of keywords, the procedure distinguishes between several conceptually different categories of expressions depending on whether the caption explicitly attributes an image to AI generation, refers only to AI-assisted editing, or only mentions artificial intelligence without clearly indicating the origin of the accompanying image.

Table A.51 summarizes the six categories comprising the disclosure dictionary. The first three categories identify captions that explicitly attribute an image to AI generation, either through direct statements, descriptive labels, or references to specific image-generation systems. A separate category identifies disclosures referring only to AI-assisted editing or enhancement, while two additional categories distinguish broader references to artificial intelligence from expressions likely to generate false-positive matches.

The dictionary produces four related disclosure measures. A *strict AI-image disclosure* is assigned whenever a caption contains either an explicit AI-generation statement, an AI visual label, or the name of an AI image-generation model. An *AI editing disclosure* identifies captions describing AI-assisted editing or enhancement. An *any AI disclosure* combines these two definitions. Finally, a *broad AI reference* identifies captions containing generic references to AI,

Table A.51: Categories used in the AI disclosure dictionary

Category	Purpose	Representative expressions
Explicit AI generation	Directly states that an image was generated using artificial intelligence	"AI-generated", "generated with AI", "created using AI", "machine-generated image"
AI visual labels	Labels the accompanying visual as AI-generated or synthetic	"AI illustration", "AI artwork", "synthetic image", "generative AI image"
Named image generators	References specific AI image-generation systems	DALL·E, Midjourney, Stable Diffusion, Adobe Firefly, Imagen, Flux, Ideogram, Leonardo AI, Runway ML
AI editing	Indicates that an existing image was modified using AI rather than fully generated	"AI-enhanced", "edited with AI", "AI retouched", "modified using AI"
Broad AI references	Generic references to artificial intelligence that may or may not describe image generation	"artificial intelligence", "generative AI", "AI model", "AI software"
Exclusion patterns	Expressions where "AI" refers to the subject of the news story rather than the image itself	"AI conference", "AI company", "AI regulation", "AI researcher", "AI policy"

Notes: This table summarizes the categories comprising the rule-based AI disclosure dictionary. The complete regular-expression dictionary used for classification is provided in the replication materials. Representative expressions are illustrative rather than exhaustive.

that do not fit the patterns described before.

The rule-based classifier is applied only to captions available in English, either because they were originally written in English or because they were successfully translated. For every figure occurrence, the procedure records the matched disclosure categories together with the specific AI image-generation systems mentioned whenever applicable. These occurrence-level disclosure classifications constitute the basis for all analyses of disclosure behavior presented in the main text.

A.8 Topic Assignment

News outlets differ substantially in how they organize their editorial content. Some publishers divide articles into conventional sections such as politics, business, sport, or culture, while others use more specialized or outlet-specific categories. To construct a common measure of article topic across the outlets in our sample, we therefore implement a two-stage procedure. First,

we recover the outlet-specific editorial themes encoded in article URL structures. Second, we harmonize these outlet-specific themes to the top-level categories of the IPTC Media Topics taxonomy.

A.8.1 Identification of Outlet-Specific Article Themes

Our first step identifies whether, and where, each news outlet encodes an editorial subject taxonomy in its article URLs. News publishers frequently organize articles using URL paths such as `domain.com/politics/article-title` or `domain.com/business/article-title`. However, this structure is not standardized across publishers. The path position containing the editorial theme can differ across outlets, and other path components may instead encode geographic locations, dates, programmes, article identifiers, or other features unrelated to an article's subject. We therefore identify the theme-bearing URL position separately for each outlet and data source.

We begin by canonicalizing outlet domains and parsing each unique article URL into its constituent path components. We retain the first four path positions and, for every dataset–outlet–position combination, characterize the empirical distribution of the values appearing in that position. Specifically, we record the share of the outlet's article URLs for which the position is present, the number of distinct values observed at the position, the frequency of each value across article URLs, and the cumulative coverage of the most frequently occurring values. These statistics help distinguish positions representing a recurring editorial taxonomy from positions containing predominantly article-specific information. For example, a position in which a relatively small set of recurring values such as `politics`, `business`, and `sport` accounts for a large share of article URLs is consistent with an editorial subject taxonomy, whereas a position containing thousands of infrequently repeated article slugs is not.

We then use GPT-4.1-mini to determine which, if any, of the first four URL path positions represents the outlet's broad editorial subject taxonomy. For each dataset–outlet combination, the model receives the structural statistics described above together with the most frequently occurring component values at each available path position. The classification task is deliberately restricted to identifying the structure of the URL rather than assigning substantive topics: the model determines whether a broad editorial subject taxonomy exists and, when it does, identifies the corresponding path position. The prompt explicitly excludes positions primarily representing geographic classifications, dates, identifiers, article slugs, programmes, technical website structure, or generic values such as `news` or `article`. When several positions appear potentially relevant, the model is instructed to select the broadest and most general subject taxonomy.

To assess the sensitivity of this classification to the amount of URL information supplied to the model, we repeat the procedure three times for every dataset–outlet combination. The three

specifications provide, respectively, the 20, 30, and 40 most frequent component values observed at each path position, together with their frequencies and structural statistics. We then use the classification receiving at least two of the three votes as the outlet-level classification.

Table A.52 summarizes the resulting classifications. The procedure covers 169 CC-News outlets containing 176,369 unique article URLs and 240 GDELT VGKG outlets containing 202,462 unique article URLs. Under the two-of-three majority classification, a broad editorial taxonomy is identified for 117 CC-News outlets (69.2%) and 158 GDELT VGKG outlets (65.8%). Although these outlets constitute approximately two thirds of the outlets in each source, they account for a larger share of the article population: 132,304 CC-News URLs (75.0%) and 154,120 GDELT VGKG URLs (76.1%). Thus, outlets for which a broad editorial taxonomy can be recovered from the URL structure account for approximately three quarters of the article URLs in both datasets.

Table A.52: Identification of editorial subject taxonomies in article URLs

	CC-News		GDELT VGKG	
	Outlets	Article URLs	Outlets	Article URLs
Panel A. Sample				
Total	169	176,369	240	202,462
Panel B. 2-of-3 majority classification				
Broad editorial taxonomy	117 (69.2%)	132,304 (75.0%)	158 (65.8%)	154,120 (76.1%)
No broad editorial taxonomy	50 (29.6%)	44,063 (25.0%)	80 (33.3%)	48,339 (23.9%)
Insufficient evidence	1 (0.6%)	1 (0.0%)	2 (0.8%)	3 (0.0%)
No 2-of-3 majority	1 (0.6%)	1 (0.0%)	0 (0.0%)	0 (0.0%)

Notes: Panel A reports the number of news outlets and unique article URLs entering the URL-taxonomy classification procedure. Panel B reports classifications based on the two-of-three majority across specifications using the 20, 30, and 40 most common URL components at each path position. "Broad editorial taxonomy" indicates that at least two specifications identified a URL path position as representing a broad editorial subject taxonomy; "No broad editorial taxonomy" indicates that at least two specifications found no such taxonomy; and "Insufficient evidence" indicates that at least two specifications judged the available URL structure insufficient to determine whether such a taxonomy exists. "No 2-of-3 majority" denotes outlets for which no classification received at least two votes. Article URLs inherit the majority classification of their outlet. Percentages, reported in parentheses, are calculated separately over outlets and article URLs within each source.

For outlets classified as having a broad editorial taxonomy, the majority-selected path position provides an outlet-specific rule for extracting the article's native editorial theme. Applying these rules to the article population produces categories such as `politics`, `business`, `technology`, and `sport`, as well as publisher-specific categories whose terminology and level of specificity

vary across outlets. For example, if the first URL component is identified as an outlet's editorial taxonomy, an article whose path begins `/politics/` receives the native theme `politics`; if the second component is identified instead, the corresponding second path component is used.

Importantly, these categories remain in their outlet-native form at this stage. We do not assume that identically or similarly named themes across different publishers necessarily correspond to the same substantive category, nor do we impose a common taxonomy during the URL-structure classification. The subsequent harmonization step maps these heterogeneous native editorial themes to a common IPTC taxonomy.

A.8.2 Harmonization to IPTC Media Topics

The outlet-native themes recovered from URL structures are not directly comparable across publishers. Different outlets frequently use different labels for substantively similar editorial categories, while others divide similar subject areas at different levels of granularity. For example, related economic coverage may appear under native labels such as `business`, `economy`, `markets`, or `money`. We therefore introduce a second classification stage that maps these heterogeneous outlet-native themes to a common substantive taxonomy.

We use the top level of the IPTC Media Topics taxonomy, a hierarchical classification system designed for news content. The top level distinguishes 17 broad subject areas: (1) *arts, culture, entertainment and media*; (2) *crime, law and justice*; (3) *disaster, accident and emergency incident*; (4) *economy, business and finance*; (5) *education*; (6) *environment*; (7) *health*; (8) *human interest*; (9) *labour*; (10) *lifestyle and leisure*; (11) *politics*; (12) *religion and belief*; (13) *science and technology*; (14) *society*; (15) *sport*; (16) *conflict, war and peace*; and (17) *weather*. We use this broader level of the hierarchy to accommodate differences in the granularity with which publishers organize their editorial content.

The harmonization begins with the outlet-specific URL position identified in the preceding step. We retain only outlets for which at least two of the three URL-structure classifications identify a broad editorial taxonomy and at least two broad-taxonomy classifications identify the same URL path position. Once this position has been established, we recover all distinct native themes observed at that position from the complete URL-component profile. The top-20, top-30, and top-40 restrictions used in the preceding classification stage therefore determine only which URL position represents the editorial taxonomy, they do not restrict the set of native themes recovered from that position.

Before harmonization to IPTC, we restrict our attention to native themes appearing in at least 1% of all article URLs associated with the corresponding outlet. Importantly, this threshold is calculated relative to the outlet's complete article corpus rather than only among URLs in which the selected taxonomy position is present. The restriction focuses the harmonization on recurring themes of publishers' editorial taxonomies and excludes rare, outlet-specific URL values from

the subsequent classification.

We then use GPT-4.1-mini to harmonize the qualifying native themes to the 17 IPTC top-level Media Topics. Classification is conducted at the outlet level. For each dataset–outlet combination, the model receives all qualifying native themes in a single prompt, together with their article frequencies and shares. Providing the complete qualifying taxonomy allows short, translated, generic, or publisher-specific labels to be interpreted in the context of the outlet's surrounding editorial categories. The model returns a separate classification for every native theme, so that each outlet–theme combination receives a single harmonized classification that can subsequently be applied to its article URLs.

The classification prompt instructs the model to assign each native theme to the single IPTC top-level category providing the closest substantive analogue. Classification is based on semantic meaning rather than lexical similarity. Where a native category spans several subjects, the model selects a single IPTC category only when one subject can reasonably be regarded as primary. We explicitly allow a valid null classification when no defensible substantive analogue exists. This includes, in particular, geographic categories such as `world` or regional names, content formats such as `video` or `podcasts`, programme or branded-section names, generic navigational labels, and categories whose substantive meaning remains ambiguous.

Table A.53 summarizes the coverage of the harmonization procedure:

Table A.53: Coverage of outlet-theme harmonization to IPTC Media Topics

	CC-News		GDELT VGKG	
	Native themes	Article URLs	Native themes	Article URLs
Eligible for harmonization	1,357	123,286	1,625	138,780
Mapped to IPTC	856 (63.1%)	67,064 (54.4%)	986 (60.7%)	67,356 (48.5%)
Valid null classification	501 (36.9%)	56,222 (45.6%)	639 (39.3%)	71,424 (51.5%)
Classification error	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Notes: The table summarizes coverage of the harmonization of outlet-native editorial themes to the 17 top-level IPTC Media Topics. Native themes are eligible for harmonization when they account for at least 1% of all unique article URLs associated with the corresponding outlet. "Mapped to IPTC" denotes eligible native themes assigned to one of the 17 substantive IPTC top-level Media Topics. "Valid null classification" denotes eligible themes for which the classifier determined that no defensible substantive IPTC analogue exists. "Classification error" denotes invalid, incomplete, or otherwise unsuccessful model outputs. Percentages in parentheses are calculated separately over eligible native themes and their associated article URLs within each source.

In CC-News, 1,357 outlet-native themes, collectively associated with 123,286 article URLs, meet the 1% eligibility threshold. Of these themes, 856 (63.1%) are successfully mapped to one of the 17 substantive IPTC categories, while 501 (36.9%) receive a valid null classification.

The corresponding figures are similar for GDELT VGKG: 1,625 native themes associated with 138,780 article URLs are eligible for harmonization, of which 986 (60.7%) are mapped to an IPTC category and 639 (39.3%) receive a valid null classification. No classification errors occur in either source.

The URL-weighted coverage is somewhat lower than the theme-level coverage, although the difference is more pronounced in GDELT VGKG. In CC-News, the 856 successfully mapped themes account for 67,064 article URLs, or 54.4% of URLs associated with eligible themes, compared with 63.1% of eligible themes. In GDELT VGKG, the 986 successfully mapped themes account for 67,356 article URLs, or 48.5% of URLs associated with eligible themes, compared with 60.7% of eligible themes. Thus, approximately three fifths of eligible outlet-native themes have a defensible substantive IPTC analogue in both sources. These mapped themes account for slightly more than half of the associated article URLs in CC-News and slightly less than half in GDELT VGKG.

Table A.54 describes the resulting composition of the article samples. The distribution is concentrated in a relatively small number of broad topics, but the composition differs across sources.

In CC-News, *sport* is the largest category, accounting for 16,295 article URLs, or 24.3% of all successfully mapped URLs. It is followed by *economy, business and finance* (20.0%), *politics* (15.3%), *arts, culture, entertainment and media* (14.1%), and *lifestyle and leisure* (10.6%). Together, these five categories account for 84.3% of successfully mapped CC-News article URLs.

The GDELT VGKG distribution differs in the relative prominence of sport. *Economy, business and finance* is the largest category, accounting for 15,323 article URLs, 22.7% of mapped URLs, followed by *arts, culture, entertainment and media* (20.1%), *politics* (17.6%), *lifestyle and leisure* (10.6%), and *sport* (9.0%). Together, these categories account for 80.0% of mapped GDELT URLs.

The resulting classification provides a common article-topic measure across outlets, countries, and data sources while remaining anchored in publishers' own editorial organization rather than inferring topics directly from article text. The two sources share several broad compositional features, while still exhibiting differences in the relative distribution of topics. These differences in topic composition indicate that subsequent comparisons of topic-specific outcomes should not be interpreted as if the two sources contain identically distributed news content.

B 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

Table A.54: Distribution of harmonized article URLs across IPTC Media Topics

	CC-News		GDELT VGKG	
IPTC Media Topic	Native themes	Article URLs	Native themes	Article URLs
Arts, culture, entertainment and media	134	9,447 (14.1%)	179	13,539 (20.1%)
Crime, law and justice	27	1,449 (2.2%)	33	1,976 (2.9%)
Disaster, accident and emergency incident	3	140 (0.2%)	4	141 (0.2%)
Economy, business and finance	146	13,439 (20.0%)	181	15,323 (22.7%)
Education	17	352 (0.5%)	16	585 (0.9%)
Environment	21	334 (0.5%)	22	617 (0.9%)
Health	39	1,179 (1.8%)	53	1,695 (2.5%)
Human interest	18	1,902 (2.8%)	21	1,197 (1.8%)
Labour	5	136 (0.2%)	5	132 (0.2%)
Lifestyle and leisure	124	7,135 (10.6%)	142	7,167 (10.6%)
Politics	98	10,263 (15.3%)	110	11,841 (17.6%)
Religion and belief	3	93 (0.1%)	3	61 (0.1%)
Science and technology	55	1,814 (2.7%)	84	3,440 (5.1%)
Society	26	2,753 (4.1%)	30	3,422 (5.1%)
Sport	132	16,295 (24.3%)	94	6,059 (9.0%)
Conflict, war and peace	1	3 (0.0%)	1	2 (0.0%)
Weather	7	330 (0.5%)	8	159 (0.2%)
Total	856	67,064 (100.0%)	986	67,356 (100.0%)

Notes: The table reports the distribution of successfully harmonized outlet-native themes and article URLs across the 17 top-level IPTC Media Topics. "Native themes" reports the number of distinct outlet-specific native themes assigned to each IPTC category. "Article URLs" reports the number of article URLs carrying those native themes. Percentages in parentheses are calculated relative to all article URLs receiving a substantive IPTC classification within the corresponding source and therefore sum to 100% within each dataset. Themes receiving a valid null classification and classification errors are excluded.

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 Figures D.2 and 5a and Table C.4.

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 D.5, Tables C.6 and C.7); (iii) **weak label effects**: labels move recognition only slightly, so they do not close the demand differences between respondents who correctly and incorrectly identified the image source (Table C.6); 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 D.3a.

B.1 Primitives

Outlet type The outlet has an unobservable editorial type

$$\theta \in \{H, L\},$$

where H denotes high editorial standards (accurate, trustworthy, professionally produced content) and L denotes lower standards. The reader holds a prior

$$\pi := () \in (0,1),$$

which is common knowledge and does not itself depend on the image condition. It reflects the reader's background belief about the outlet before exposure to any article.

Technology menu In each production period the outlet chooses a technology $a \in \{\text{AI}, \text{real}\}$ for the article image. Technologies differ along two dimensions: variable cost and recognisability.

Definition B.1 (Technology pairs)**.** Let κ_R denote the variable cost of a real wire-service photograph and κ_{AI} the cost of an AI-generated image, with $\kappa_{\text{AI}} < \kappa_R$. The cost saving is $\delta := \kappa_R - \kappa_{\text{AI}} > 0$. Let $\rho \in [\rho^{\min}, \rho^{\max}]$ be a *photorealism parameter* chosen jointly with a: higher ρ (more realistic AI) lowers the probability that the reader classifies the image as synthetic.

The photorealism parameter ρ captures the Experiment 2 manipulation, where hard-to-detect and easy-to-detect AI images are randomly assigned.

Disclosure regime The outlet additionally chooses a binary disclosure label $\ell \in \{0,1\}$, where $\ell = 1$ means a source cue is displayed (a wire-service credit for real photos; an AI-disclosure tag for synthetic ones). The pair (a,ℓ) fully characterises a treatment arm.

B.2 The Reader's Information Problem

Image as a noisy signal After reading the article the reader observes the image but cannot directly observe (a,ℓ). Instead, she forms a binary recognition judgment

$$r \in \{0,1\}, \qquad r = 1 \Leftrightarrow \text{“this looks AI-generated to me.”}$$

Conditional on the outlet's type and the label status, recognition follows

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

[Diagnosticity] For both label conditions,

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

The intuition for Assumption B.2 is that a low-standard outlet is more likely to put the reader in a situation where the image reads as synthetic, either because it actually used AI, used it carelessly, or provided weak contextual cues. Seeing $r = 1$ is therefore relatively more likely under .

Two qualifications apply. Empirically, the assumption holds in Experiment 2's unlabeled arms ($p_L = 0.441 > p_H = 0.367$) but is marginally violated in Experiment 1's ($0.458 < 0.472$), suggesting that for highly photorealistic images the diagnostic content of the recognition judgment is weak (see Section B.13.4). Conceptually, the assumption embeds an equilibrium correlation between AI use and outlet type that our experimental design deliberately breaks by holding content quality constant; the welfare implications of this tension are treated in Section B.15.6.

[Label effects] Labels shift recognition in opposite directions depending on origin:

$$p_{H,1} < p_{H,0} \approx p_{L,0} < p_{L,1}.$$

That is, a wire-service credit slightly reduces false AI attributions for real photos, while an AI disclosure slightly increases correct detections for synthetic ones. The shifts are modest consistent with the empirical finding that adding a label moves recognition rates by only four to six percentage points.

Bayesian belief updating Given the signal (r, ℓ), the reader updates her belief about the outlet's type using Bayes' rule:

$$\mu(1,\ell) := (\mid r=1,\ell) = \frac{\pi\, p_{H\ell}}{\pi\, p_{H\ell} + (1-\pi)\, p_{L\ell}}, \tag{5}$$

$$\mu(0,\ell) := (\mid r=0,\ell) = \frac{\pi(1-p_{H\ell})}{\pi(1-p_{H\ell}) + (1-\pi)(1-p_{L\ell})}. \tag{6}$$

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

$$\mu(1,\ell) \;<\; \pi \;<\; \mu(0,\ell) \quad \textit{for } \ell \in \{0,1\}.$$

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

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

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

Lemma B.2 is the core insight: the reader rationally punishes the outlet whenever the image reads as AI-generated, because only a sloppier outlet would consistently produce that impression—regardless of whether the image is actually synthetic. Even readers exposed to an authentic wire-service photo penalise the outlet whenever they mistakenly classify it as AI.

B.3 Demand Decision

Payoffs The reader has an outside option normalised to zero. Subscribing to the newsletter yields expected utility

$$U(d=1 \mid r,\ell) \;=\; v_H + \big(1-\big)\, v_L - c,$$

where $v_H > v_L \geq 0$ are the values of future newsletters from an H- and L-type outlet respectively, and $c \geq 0$ is an idiosyncratic subscription cost drawn from a continuous CDF F on $\mathbb{R}_+$, independent of all signals. The cost c captures the heterogeneous opportunity cost of inbox attention (time, effort, signal-to-noise preferences) that generates variation in sign-up rates across readers.

Subscription rule The reader subscribes if and only if $c \leq (r,\ell)$, where

$$(r,\ell) \;:=\; (v_H - v_L) + v_L \tag{7}$$

is the willingness-to-pay threshold. Since is strictly increasing in and F is non-decreasing with positive density, we define the demand probability conditional on signal (r,ℓ) as

$$D(r,\ell) \;:=\; F((r,\ell))\,. \tag{8}$$

Proposition B.3 (Recognition penalty in demand)**.** *For each label condition $\ell \in \{0,1\}$,*

$$D(0,\ell) \;>\; D(1,\ell).$$

Proof. Lemma B.2 gives $\mu(0,\ell) > \mu(1,\ell)$. Since is strictly increasing in μ (equation (7)) and F has positive density in a neighbourhood of , it follows that $D(0,\ell) = F((0,\ell)) > F((1,\ell)) = D(1,\ell)$. □

Proposition B.3 corresponds to the ∼10 percentage-point demand gap documented in both experiments: readers who correctly identify AI images subscribe less; readers who do not subscribe more.

B.4 From Signals to Treatment Arms

Within treatment arm $t \in \{$unlabelled real, labelled real, unlabelled AI, labelled AI$\}$, the fraction of readers reporting $r = 1$ is the arm-specific attribution rate α_t. Average demand in arm t is therefore a mixture:

$$_t = \; \alpha_t\, D(1,\ell_t) + (1-\alpha_t)\, D(0,\ell_t), \tag{9}$$

where $\ell_t \in \{0,1\}$ is the label status of arm t.

Proposition B.4 (Aggregate null)**.** *Suppose $\alpha_t \in [\underline{\alpha},\bar{\alpha}]$ with $\bar{\alpha}-\underline{\alpha}$ small, and that the demand gap $D(0,\ell)-D(1,\ell)$ is approximately equal across arms. Then*

$$_t \;\approx\; \alpha\, D(1,\ell) + (1-\alpha)\, D(0,\ell) \quad \forall t,$$

and average demand is nearly constant across treatment arms.

Proof sketch. Substituting the empirical attribution rates $\alpha_t \in \{0.472, 0.429, 0.458, 0.516\}$ Table B.2 into (9), and using the fact that $D(0,\ell)-D(1,\ell)$ is approximately constant across arms (Proposition B.3), the weighted sum $_t$ varies by at most $(\bar{\alpha}-\underline{\alpha})\cdot[D(0)-D(1)]$. With $\bar{\alpha}-\underline{\alpha} \approx 0.09$ and a demand gap of 0.10, the maximum spread across arms is ≈ 0.009, consistent with the narrow 36.5–39.3% band observed in Experiment 1. □

Remark B.5. Proposition B.4 rationalises the aggregate null without assuming that AI images are irrelevant to readers. The null arises from symmetric cancellation: every arm puts roughly half the sample into the high-demand state ($r = 0$) and half into the low-demand state ($r = 1$). The

mechanism is analogous to Chopra et al. (2024), where fact-checking raises perceived accuracy among some readers while confirming priors for others, leaving average demand approximately unchanged.

B.5 The Outlet's Technology Choice

Per-period profit In each period the outlet maximises profit Π over the technology choice (a, ρ, ℓ):

$$\Pi(a, \rho, \ell) \;=\; (a, \rho, \ell) \cdot m - \kappa(a) - \phi(\ell), \tag{10}$$

where $m > 0$ is the margin per subscriber (advertising or subscription revenue), $\kappa(a)$ is the variable image-production cost, and $\phi(\ell) \geq 0$ is the disclosure cost (legal, operational, or reputational). Since Proposition B.4 implies is approximately invariant across arms, the outlet always weakly prefers AI adoption on direct cost grounds:

$$\Pi(\text{AI}, \rho, \ell) - \Pi(\text{real}, \rho, \ell) \;\approx\; \delta - \Delta\phi(\ell) \;\geq\; 0 \quad \text{when } \delta \geq \Delta\phi(\ell).$$

Reputational cost of recognition This comparison holds only for *average* demand. Define the recognition rate under AI adoption as $\alpha_{\text{AI}}(\rho, \ell)$, which is decreasing in photorealism ρ and increasing in label presence ℓ (by Assumption B.2). The demand shortfall from recognised AI, relative to the real-photo counterfactual, is

$$\Delta D^{\text{rep}} \;=\; \alpha_{\text{AI}}(\rho, \ell) \left[D(0, \ell_R) - D(1, \ell_{\text{AI}}) \right], \tag{11}$$

where the bracketed term is the demand gap guaranteed by Proposition B.3.

The outlet therefore faces a cost–recognition trade-off. Switching to AI reduces κ but inflates α_{AI} and erodes demand among recognising readers. The photorealism parameter ρ is the internal instrument: increasing ρ lowers recognition (and thus the reputational cost) at the expense of higher AI generation quality costs. The interior optimum satisfies

$$\frac{\partial \kappa_{\text{AI}}}{\partial \rho} \;=\; -m \cdot \frac{\partial \alpha_{\text{AI}}}{\partial \rho} \cdot \left[D(0, \ell) - D(1, \ell) \right], \tag{12}$$

equating the marginal cost of higher image quality to the marginal demand gain from reduced recognition.

B.6 The Role of Disclosure

Label as a belief shifter A disclosure label $\ell = 1$ operates through two channels. The *direct channel* raises $p_{L,1}$ (AI images are more often correctly identified when labelled) and lowers

$p_{H,1}$ (real photos are somewhat less likely to be mistaken for AI). The *indirect channel* shifts the mixture weights in $_t$ via equation (9) but does not change the height of the gap $D(0,\ell) - D(1,\ell)$.

Proposition B.6 (Label cannot close the recognition gap)**.** *Under Assumption B.2 (modest label effects), even with* $\ell = 1$*,*

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

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

Proof. The gap $D(0,\ell) - D(1,\ell) = F((0,\ell)) - F((1,\ell))$ depends on $\mu(0,\ell) - \mu(1,\ell)$, which in turn depends on $p_{\theta\ell}$ through equations (5)–(6). Assumption B.2 restricts labels to shift $p_{\theta\ell}$ by at most a few percentage points. For the gap to close entirely, we would need $\mu(0,\ell) \to \mu(1,\ell)$, which requires either $p_{H\ell} \to p_{L\ell}$ (perfect recognition) or F to be degenerate at a common threshold. Neither condition holds in the data. □

B.7 Transparency bonus for realistic AI imagery

A nuanced result emerges for hard-to-detect AI images combined with a label. When the AI image is photorealistic (ρ high), the baseline recognition rate α_{AI} is low. In this case, the label serves a different function: it preempts the inference "this outlet uses AI carelessly" and replaces it with "this outlet is transparent about AI use." Formally, among readers who would have remained sceptical without the label, the label shifts the prior on the outlet's type upward, partially offsetting the recognition penalty. This is consistent with the labelled hard-AI arm of Experiment 2, where the demand penalty among correctly recognising readers is near zero.

B.8 Equilibrium Characterisation

The model has the following sequential structure with no backward induction loop:

1. **Nature** draws $\theta \in \{H, L\}$ with probability $(\pi, 1-\pi)$.
2. **Outlet** observes θ and chooses (a, ρ, ℓ) to maximise $\Pi(a, \rho, \ell)$.
3. **Reader** observes (image, ℓ), forms recognition judgment r, updates to , draws private cost $c \sim F$, and subscribes iff $c \leq (r, \ell)$.

Because θ is drawn before the outlet acts and the reader does not observe θ directly, the model is a Bayesian signalling structure. Crucially, the reader's demand decision does not feed back into θ within the period: quality is fixed, not endogenously sustained by readership. This removes the circularity present in Klein–Leffler reputation cycles.

Table B.1: Comparative statics

Parameter change	Effect on $_t$	Effect on AI adoption incentive
Higher photorealism $\rho \uparrow$	Ambiguous (shifts α_t down, raises $D(0,\ell)$ path weight)	Increases AI adoption
Larger cost saving $\delta \uparrow$	None (demand approx. invariant)	Increases AI adoption
Stronger label effects (higher literacy)	Small increase for AI arms; small decrease for real arms	Reduces AI adoption if reputational cost becomes salient
Higher reader prior $\pi \uparrow$	Demand rises uniformly; gap $D(0,\ell) - D(1,\ell)$ shrinks	Reduces reputational cost of recognition
Larger value spread $v_H - v_L \uparrow$	Gap $D(0,\ell) - D(1,\ell)$ widens	Increases reputational cost of recognition

Comparative statics Table B.1 summarises the key comparative statics of the model.

B.9 Extension: Heterogeneous Reader Types

The baseline model treats all readers symmetrically. A natural extension stratifies readers by AI literacy $\lambda \in [0,1]$, where λ captures the probability that a reader correctly identifies the image type regardless of label. Under this extension:

- High-λ readers always recover r correctly, so their beliefs coincide with the true image origin.
- Low-λ readers behave as in the baseline: their r is noisy and governed by $p_{\theta\ell}$.

Aggregate demand becomes

$$D_t^{\lambda} \;=\; \lambda\, D(r^*,\ell) + (1-\lambda)\, D_t, \tag{13}$$

where r^* is the correct recognition outcome for that arm.

Corollary B.7 (Dissolution of the aggregate null)**.** *As the population becomes more AI-literate ($\lambda \to 1$), the aggregate null dissolves: recognising AI imagery correctly now consistently reduces demand, eliminating the offsetting cancellation mechanism.*

Corollary B.7 predicts that, as AI detection tools diffuse and media-literacy interventions take effect, the empirical null will shift toward a monotone negative average effect of AI adoption on demand. This is a testable comparative static for future work.

B.10 Key Predictions

The framework generates five testable predictions, all consistent with the empirical findings.

1. **Aggregate null.** Average newsletter demand is approximately constant across treatment arms, because recognition rates cluster near 50% in every arm, generating offsetting positive and negative demand shifts that cancel in the population mixture.

2. **Belief-driven heterogeneity.** Conditional on believing the image is real ($r = 0$), demand is uniformly higher; conditional on believing it is AI ($r = 1$), demand is uniformly lower—regardless of whether that belief is correct.

3. **Label modesty.** Labels move average demand little: they shift recognition rates only modestly (Assumption B.2), and the demand gap $D(0,\ell) - D(1,\ell)$ is determined by the prior π and the value spread $v_H - v_L$, not by the label per se.

4. **Transparency bonus.** For hard-to-detect AI images, labelling can partially protect demand among recognising readers by reframing the inference from "careless AI use" to "transparent AI use." This is a second-order effect and depends on reader sophistication.

5. **Reputational cost of recognised AI.** Even when the AI image is of objectively high quality (i.e., the outlet's true type is H), recognition of AI use triggers a downward belief revision, because readers cannot observe θ and update on the signal $r = 1$ as if AI use correlates with lower standards. The outlet bears a reputational cost that is independent of actual quality.

B.11 Quantifying the Dissolution of the Aggregate Null

Corollary B.7 establishes qualitatively that as AI literacy $\lambda \to 1$, the aggregate null dissolves and average demand falls monotonically with AI adoption. This section makes the threshold precise using the mixture equation (9) and the Experiment 2 moments in Table B.4. All quantities in this and the following sections are computed from Experiment 2; no Experiment 1 moments enter the structural exercises.

Belief-conditional demand rates. Unlike a decomposition from aggregate moments, Experiment 2 permits direct estimation of the belief-conditional demand rates as sample conditional

means:

$$D^- = 0.322 \quad \text{(believe AI)}, \qquad D^+ = 0.384 \quad \text{(believe real)}, \tag{14}$$

a gap of $D^+ - D^- = 6.2$ percentage points (Table B.4). These are the structural demand rates the outlet faces, separated entirely by audiences' beliefs rather than by the objective origin of the image.

Parameterising AI literacy. Let $\lambda \in [0,1]$ index joint improvement in AI-image detection and real-image authentication, with $\lambda = 0$ the current state and $\lambda = 1$ full information. Anchoring on the Experiment 2 unlabeled-arm attribution rates ($p_L = 0.441$ for the photorealistic hard-AI image an outlet would plausibly deploy; $p_H = 0.367$ for authentic photographs), we model recognition as

$$\alpha_{\text{AI}}(\lambda) = 0.441 + 0.559\,\lambda, \tag{15}$$

$$\alpha_{\text{REAL}}(\lambda) = 0.367\,(1-\lambda). \tag{16}$$

Equation (15) captures rising correct recognition of synthetic images; equation (16) captures the declining false-positive rate for authentic photographs. The boundary conditions are $\alpha_{\text{AI}}(0) = 0.441$, $\alpha_{\text{AI}}(1) = 1$ (perfect detection of AI), $\alpha_{\text{REAL}}(0) = 0.367$, and $\alpha_{\text{REAL}}(1) = 0$ (no false positives under full information). This parameterisation coincides with the one used in the welfare analysis (equation (41)), so all exercises share a single technology path.

Closed-form demand gap. With $\bar{D}(\alpha) = D^+ - (D^+ - D^-)\,\alpha$, substituting (15)–(16):

$$\bar{D}_{\text{AI}}(\lambda) = 0.356 - 0.034\,\lambda, \tag{17}$$

$$\bar{D}_{\text{REAL}}(\lambda) = 0.361 + 0.023\,\lambda, \tag{18}$$

$$\text{Gap}(\lambda) \equiv \bar{D}_{\text{AI}}(\lambda) - \bar{D}_{\text{REAL}}(\lambda) = -0.005 - 0.057\,\lambda. \tag{19}$$

The gap is linear in λ, mildly negative at the current state (because $p_L > p_H$: photorealistic AI is already attributed to AI somewhat more often than authentic photographs), and reaches -6.2 percentage points under full information. As an out-of-sample check, the model-implied current gap of -0.5 p.p. closely matches the observed -0.6 p.p. difference between the unlabeled hard-AI and unlabeled control arms (Table C.15), a moment not used in the calculation.

Critical literacy threshold. Statistical detectability depends on the design. At Experiment 2's total sample size ($N = 4{,}196$) split evenly between AI and real conditions, the minimum detectable effect is $\text{MDE} = 2.8\sqrt{2\bar{p}(1-\bar{p})/n} \approx 4.1$ p.p. ($\bar{p} = 0.35$, $n = 2{,}098$; 80% power, two-sided $\alpha = 0.05$). At the actual per-arm sizes ($n \approx 700$), pairwise arm comparisons have an MDE of 7.1 p.p., which exceeds even the full-information gap; the arm-level null is therefore robust

not only because beliefs offset but because the belief-conditional gap is itself modest. Setting $0.005 + 0.057\,\hat{\lambda} = 0.041$:

$$\hat{\lambda} \approx 0.64. \tag{20}$$

At $\hat{\lambda}$: $\alpha_{\text{AI}} = 79.9\%$ (up from 44.1% today) and $\alpha_{\text{REAL}} = 13.3\%$ (down from 36.7%). Table B.2 tabulates the gap at selected values of λ.

Table B.2: Demand Gap as a Function of AI Literacy λ (Experiment 2 moments)

λ	α_{AI}	α_{REAL}	$\bar{D}_{\text{AI}}$	Gap (pp)	Null holds
0.00 (today)	44%	37%	35.6%	−0.5	Yes
0.25	58%	28%	34.8%	−1.9	Yes
0.50	72%	18%	33.9%	−3.3	Yes
0.64	**80%**	**13%**	**33.4%**	**−4.1**	**Borderline**
0.80	89%	7%	32.9%	−5.0	No
1.00	100%	0%	32.2%	−6.2	No

Notes: MDE = 4.1 pp for an equal-split design at $N = 4{,}196$ (80% power, two-sided $\alpha = 0.05$). $D^- = 32.2\%$ and $D^+ = 38.4\%$ are the Experiment 2 belief-conditional sample means (Table B.4). $\hat{\lambda} \approx 0.64$ from equation (20). α_{AI} and α_{REAL} at each λ from equations (15)–(16).

Two observations follow from equation (19) and Table B.2. First, reducing false positives for real photographs matters as much as improving AI detection: the currently high false-positive rate of 37% keeps control-group demand artificially depressed, compressing the gap. Authentication tools for genuine photojournalism such as the C2PA Content Credentials standard therefore contribute as much to dissolving the null as AI-detection classifiers do. Second, even at $\lambda = 1$ the demand gap is bounded by $D^+ - D^- = 6.2$ percentage points, placing a finite ceiling on aggregate harm through the demand channel alone. The causal quality penalty of 0.348 SD (Table C.17, Section 4.5) operates through a separate reputational channel and is likely to dominate in the long run.

B.12 Outlet Breakeven Cost of AI Visual Adoption

Drawing on the outlet's profit function (10), AI adoption is profitable when the per-image cost saving δ exceeds the recognition-driven demand shortfall. Using $\text{Gap}(\lambda)$ from equation (19), this condition is

$$\delta > -\text{Gap}(\lambda) \cdot m \cdot R_{\text{eff}} = \big(0.005 + 0.057\,\lambda\big) \cdot m \cdot R_{\text{eff}}, \tag{21}$$

where m is the lifetime revenue per newsletter subscriber and R_{eff} is the effective audience per article (readers who encounter the sign-up prompt). Rearranging, the breakeven literacy level is

$$\lambda^*(\delta, m, R_{\text{eff}}) = \frac{\delta}{0.057 \cdot m \cdot R_{\text{eff}}} - 0.08. \tag{22}$$

For $\lambda < \lambda^*$ AI adoption is profitable; for $\lambda > \lambda^*$ the engagement loss outweighs the cost saving. Because the current gap is already negative, $\lambda^* < 0$ is possible: setting $\lambda^* = 0$ defines the critical audience scale

$$R^* = \frac{\delta}{0.005 \cdot m}, \tag{23}$$

above which AI imagery is unprofitable at *current* detection rates.

Anchoring δ: wire-service licensing costs. Editorial image licensing fees from major wire agencies range from approximately £50 for standard regional-agency use to £500 or more for exclusive images from Reuters, AP, or Getty (Associated Press, 2024; Getty Images, 2024). Survey evidence consistently identifies cost reduction including the elimination of per-image licensing fees as the primary operational motivation for adopting AI visuals in newsrooms (Beckett and Yaseen, 2023; Diakopoulos et al., 2024). We use £150 as a central estimate, consistent with mid-tier editorial licensing for a digital-first outlet such as *The Conversation*. All qualitative conclusions hold across the full plausible range £50–£500.

Anchoring m: newsletter subscriber lifetime value. For an advertisement-supported outlet, annual direct revenue per subscriber is $(\text{CPM}/1000) \times$ open rate $\times$ issues p.a. UK email open rates average 35–42% and newsletter CPMs range from £20 to £35 (Interactive Advertising Bureau UK, 2024; Litmus Software, Inc., 2024). At 40 weekly issues and a CPM of £25, direct advertising revenue is approximately £4 per subscriber per year. Including the probability of upselling to paid subscriptions and the brand-loyalty value of retained readers, industry benchmarks suggest a lifetime value of £5–£25 (Interactive Advertising Bureau UK, 2024; Reuters Institute for the Study of Journalism, 2024). We use $m = 15$ as a central estimate.

Anchoring R_{eff}: effective audience per article. *The Conversation UK* attracts approximately three million unique monthly visitors across around 200 articles per month, implying roughly 15,000 readers per article (The Conversation, 2024). Not every reader encounters the newsletter sign-up prompt; assuming a 5–20% exposure rate gives $R_{\text{eff}} \in [750, 3{,}000]$ as a central range for this outlet. We tabulate results across $R_{\text{eff}} \in [100, 10{,}000]$ to cover micro-outlets and major national publishers alike.

Results. Substituting $\delta = 150$ and $m = 15$ into (22) yields $\lambda^* \approx 175/R_{\text{eff}} - 0.08$ and, from (23), a critical scale of $R^* \approx 2{,}200$ effective readers per article. Table B.3 shows how the breakeven threshold shifts with outlet scale.

Interpretation. Two features of Table B.3 are noteworthy. First, λ^* is steeply declining in R_{eff} and crosses zero at $R^* \approx 2{,}200$: because photorealistic AI already carries a small recognition

Table B.3: Breakeven AI Literacy Threshold by Outlet Scale ($\delta = 150$, $m = 15$; Experiment 2 moments)

Outlet scale (R_{eff})	λ^*	α_{AI} at λ^*	Engagement loss today	Assessment
100 (micro)	> 1	—	£7	Profitable at any λ
500 (small)	0.27	59%	£34	Profitable, moderate margin
1,000 (mid)	0.09	49%	£68	Narrow margin
2,000	0.01	45%	£137	At breakeven now
5,000 (large)	< 0	—	£342	Unprofitable now
10,000 (major)	< 0	—	£684	Unprofitable now

Notes: λ^* from equation (22), truncated to $[0, 1]$; $\lambda^* > 1$ means the engagement loss never reaches δ (for $R_{\text{eff}} = 100$, the loss at full information is $0.062 \cdot m \cdot R_{\text{eff}} = 92 < 150$), and $\lambda^* < 0$ means the loss exceeds δ already at current detection rates. "Engagement loss today" equals $0.005 \cdot m \cdot R_{\text{eff}}$, the current-state shortfall from equation (21) at $\lambda = 0$. α_{AI} at λ^* from equation (15). Conclusions are robust across $\delta \in [50, 500]$ and $m \in [5, 25]$; the critical scale $R^* = \delta/(0.005m)$ ranges from roughly 400 to 14,600 across this grid.

penalty at today's detection rates, any outlet with more than roughly 2,200 effective readers per article loses money on AI imagery *now*, before any improvement in AI literacy. Second, the asymmetry is structural: the cost saving δ is a fixed per-image amount, while the engagement loss $(0.005 + 0.057\lambda) \cdot m \cdot R_{\text{eff}}$ scales linearly with audience size. AI imagery is therefore a rational editorial choice for micro-outlets and an increasingly poor one for any outlet with meaningful traffic, independently of disclosure norms or editorial standards. An observable implication follows: large outlets should retreat from AI visuals substantially earlier than small outlets in response to rising AI-detection capabilities, not because of greater ethical sensitivity, but because the cost-benefit calculation turns negative sooner. Finally, equation (21) captures only the short-run demand channel. The causal quality penalty of 0.348 SD per unit of recognition (Table C.17) operates through a separate reputational channel: even where $\lambda < \lambda^*$ and short-run demand is unaffected, repeated AI use may compound credibility losses over time in ways a single-article design cannot capture.

B.13 Structural Parameter Recovery

This section recovers the parameters of the model from the experimental data. We first list the parameters and the inputs taken directly from the data (Section B.13.1), then derive the posterior beliefs (Section B.13.2) and the moment conditions linking them to observed behaviour (Section B.13.3), and finally solve for the parameters (Section B.13.4).

B.13.1 Parameters and Identification

The model has four structural parameters.

- $\pi \in (0,1)$: the reader's prior probability that the outlet is of high editorial type H.
- $\Delta v \equiv (v_H - v_L)/c_{\max} \in (0,1]$: the normalized value spread between newsletters from a high- and a low-type outlet, where $c_{\max}$ is the upper bound of the idiosyncratic subscription-cost distribution F.
- $v_L/c_{\max} \in [0,1)$: the normalized floor demand, equal to the subscription rate when the outlet is known to be of low type.
- $b \in \mathbb{R}$: the slope of the quality index on the posterior belief μ, measuring how strongly perceived quality responds to inferences about outlet type.

Two further objects are not estimated but read directly off the manipulation-check data:

$$p_H = \Pr(r=1 \mid \text{real image, no label}) = 0.367, \qquad p_L = \Pr(r=1 \mid \text{hard AI, no label}) = 0.441,$$

where $r=1$ denotes a reader's judgment that the image is AI-generated. They come from the control arm and the unlabeled hard-AI arm of Experiment 2, respectively. p_H is the false-positive rate (authentic photographs judged to be AI) and p_L the true-positive rate for the photorealistic AI images an outlet would plausibly deploy.

B.13.2 Bayesian Updating

Given (p_H, p_L, π), Bayes' rule gives the reader's posterior belief that the outlet is of high type, conditional on the recognition judgment r:

$$\mu^- \equiv \mu(r=1) = \frac{\pi\, p_H}{\pi\, p_H + (1-\pi)\, p_L}, \tag{24}$$

$$\mu^+ \equiv \mu(r=0) = \frac{\pi\,(1-p_H)}{\pi\,(1-p_H) + (1-\pi)\,(1-p_L)}. \tag{25}$$

Throughout, the superscript $-$ refers to readers who believe the image is AI and $+$ to readers who believe it is real. Since $p_L > p_H$ (Assumption B.2), Lemma B.2 gives $\mu^- < \pi < \mu^+$: recognising the image as AI is bad news about outlet quality, and not recognising it is good news.

B.13.3 Moment Conditions

With subscription costs distributed $c \sim \text{Uniform}[0, c_{\max}]$, the demand function (equation (8)) becomes linear in the posterior:

$$D(r) = \Delta v \cdot \mu(r) + v_L/c_{\max}.$$

The quality index is standardized to have mean zero in the full sample, so it depends on how far a reader's posterior lies from the sample average:

$$Q(r) = b \cdot \big(\mu(r) - \bar{\mu}\big), \qquad \bar{\mu} \equiv \bar{\alpha}\,\mu^{-} + (1-\bar{\alpha})\,\mu^{+},$$

where $\bar{\alpha} = 0.561$ is the mean attribution rate.

Matching these model predictions to their sample counterparts gives four moment conditions:

$$g_1(\theta) = D^{-} - \big(\Delta v \cdot \mu^{-}(\pi) + v_L\big), \tag{26}$$
$$g_2(\theta) = D^{+} - \big(\Delta v \cdot \mu^{+}(\pi) + v_L\big), \tag{27}$$
$$g_3(\theta) = Q^{-} - b \cdot \big(\mu^{-}(\pi) - \bar{\mu}(\pi)\big), \tag{28}$$
$$g_4(\theta) = Q^{+} - b \cdot \big(\mu^{+}(\pi) - \bar{\mu}(\pi)\big), \tag{29}$$

where $\theta = (\pi, \Delta v, v_L, b)$, $D^{\pm}$ and $Q^{\pm}$ are the sample means of demand and quality in each recognition group, and $\mu^{\pm}(\pi)$ are the posteriors in equations (24)–(25). Table B.4 reports the empirical values.

Table B.4: Empirical Moments Used in Structural Estimation

Moment	Symbol	Value
Newsletter demand | believe AI	D^{-}	0.3219
Newsletter demand | believe real	D^{+}	0.3835
Quality z-index | believe AI	Q^{-}	−0.2351
Quality z-index | believe real	Q^{+}	0.3008
Mean attribution rate	$\bar{\alpha}$	0.5612
False-positive rate (real images)	p_H	0.3670
True-positive rate (hard AI, no label)	p_L	0.4411
Demand gap $D^{+} - D^{-}$		0.0616
Quality gap $Q^{+} - Q^{-}$		0.5359

Notes: D^{-} and Q^{-} are sample means among readers who judge the image to be AI ($r = 1$); D^{+} and Q^{+} among readers who judge it to be real ($r = 0$). Experiment 2, $N = 4{,}196$. p_H and p_L are attribution rates in the control and unlabeled hard-AI arms.

B.13.4 Calibration Strategy and Parameter Estimates

The system (26)–(29) has four equations in four unknowns, but only three of the equations are independent. The quality index is standardized on the same sample, with the same attribution weights, used to construct $\bar{\mu}$, so the two quality conditions g_3 and g_4 carry a single restriction: whenever one holds, so does the other. This leaves one degree of freedom, which we treat as the

prior π.

For a given π, the two demand conditions (26)–(27) identify Δv and v_L in closed form:

$$\Delta v(\pi) = \frac{D^+ - D^-}{\mu^+(\pi) - \mu^-(\pi)}, \tag{30}$$

$$v_L(\pi) = D^- - \Delta v(\pi) \cdot \mu^-(\pi), \tag{31}$$

and the quality restriction identifies b:

$$b(\pi) = \frac{Q^+ - Q^-}{\mu^+(\pi) - \mu^-(\pi)}. \tag{32}$$

Both slopes have the same form: an observed gap between the two recognition groups divided by the model's posterior gap.

It remains to fix π. We impose the *zero-floor condition* $v_L = 0$, under which a newsletter from a known low-type outlet attracts no marginal subscribers. Substituting into (31) gives

$$v_L(\pi) = 0 \implies D^- = \Delta v(\pi) \cdot \mu^-(\pi), \tag{33}$$

a scalar equation in π that we solve numerically by Brent's method on the interval $(0.01, 0.49)$.

Remark B.8. The zero-floor condition is a normalization, not a behavioural restriction. It sets the level of π but leaves the identified gaps $D^+ - D^-$ and $Q^+ - Q^-$ unchanged, and hence leaves all welfare *changes* unaffected. All qualitative conclusions hold for $v_L \in [0, 0.07]$, the range consistent with demand remaining in $[0, 1]$, which corresponds to $\pi \in [0.30, 0.58]$.

The recovery is therefore exact rather than an estimation: the parameters solve the moment conditions in closed form, and no over-identifying restrictions remain to be tested. The model does imply two further restrictions, within-group homogeneity and cross-equation proportionality, which play no role in recovering the parameters. We use them as specification checks in Section B.14.

Table B.5 reports the estimates. The prior $\hat{\pi} = 0.439$ means that, before seeing the image, a reader assigns a 44% probability to the outlet being of high type. The value spread $\widehat{\Delta v} = 0.816$ is large: the difference in value between a high- and a low-type newsletter amounts to more than 80% of the maximum subscription cost, so the quality difference between outlet types matters substantially to readers.

The quality sensitivity $\hat{b} = 7.10$ should not be read literally. It is large for a mechanical reason. Because p_H and p_L are close (0.367 versus 0.441), recognition is only weak evidence about the outlet, and the posteriors barely move: $\mu^+ - \mu^- = 0.075$. Equation (32) divides the observed 0.54 SD quality gap by this small posterior gap, which forces an extreme slope. Taken at face value, a reader who became certain the outlet was low type would lower perceived quality

Table B.5: Structural Parameter Estimates

Parameter	Interpretation	Estimate	Bootstrap SE
π	Prior Pr(high-type outlet)	0.439	0.105
Δv	Normalized value spread $(v_H - v_L)/c_{\max}$	0.816	0.165
v_L	Floor demand (calibrated)	0.000	—
b	Quality sensitivity per unit μ	7.096	0.573
μ^-	Posterior | believe AI	0.395	—
μ^+	Posterior | believe real	0.470	—
$\mu^+ - \mu^-$	Posterior gap	0.075	—

Notes: Bootstrap standard errors based on 500 resamples with replacement. $v_L = 0$ by calibration, so no standard error is reported. $\mu^{\pm}$ follow from the estimated π via equations (24)–(25) and have no independent standard errors.

by seven standard deviations.

A more natural reading is behavioural. Respondents do not perform the small Bayesian update the model prescribes; they react categorically to the belief "this image is AI," applying a roughly fixed penalty however weakly diagnostic the signal is. Two pieces of evidence support this reading. First, the diagnosticity assumption (Assumption B.2) holds in Experiment 2's unlabeled arms ($p_L = 0.441 > p_H = 0.367$) but fails marginally in Experiment 1's (45.8% versus 47.2%), where the photorealistic AI image was attributed to AI slightly *less* often than the authentic photograph. This is inconsistent with recognition being an informative signal of type, but consistent with a categorical penalty. Second, the quality penalty among recognising readers is essentially the same whether the stimulus was easy or hard to detect (Table C.16), whereas Bayesian updating predicts larger revisions for more diagnostic signals.

We therefore treat the Bayesian structure as a tractable reduced-form representation of belief-conditional behaviour. For this reason, every object in the welfare analysis of Section B.15 depends only on the directly estimated $D^{\pm}$, $Q^{\pm}$, and arm-level attribution rates, and not on π, $\mu^{\pm}$, or b.

B.14 Model Restriction Tests

The model implies two restrictions that are neither imposed nor used in recovering the parameters in Section B.13. We test them here as checks on the belief-conditioning structure. They are not over-identifying restrictions in the GMM sense, since the parameter recovery has none.

B.14.1 Within-Group Homogeneity

The model's central testable restriction is that a reader's demand and quality perceptions depend only on her recognition judgment r, and not on which treatment arm led her to it. A reader who believes the image is real should behave the same whether she was shown an authentic photograph or an AI image she failed to recognise. Formally,

$$\mathbb{E}\big[D_i \mid r_i = r,\, t_i = t\big] = D(r) \qquad \forall\, t \in \{\text{C, LC, HAI, LHAI, EAI, LEAI}\}, \tag{34}$$

and analogously for Q_i. We test this with F-tests of equal means across the six arms within each recognition group.

Table B.6: Within-Group Homogeneity F-Tests

Outcome	Recognition group	F-statistic	df_1	df_2	p-value
Newsletter demand	Believe AI ($r = 1$)	2.102	5	2,349	0.062
Newsletter demand	Believe real ($r = 0$)	3.404	5	1,835	**0.005**
Quality z-index	Believe AI ($r = 1$)	2.605	5	2,349	**0.023**
Quality z-index	Believe real ($r = 0$)	0.769	5	1,835	0.572

Notes: F-tests of equal group means across the six treatment arms within each recognition group. The null hypothesis is equation (34). Bold p-values indicate rejection at the 5% level.

Table B.6 shows that homogeneity holds only partly. It is not rejected for quality among believe-real readers ($p = 0.572$), nor, at the 5% level, for demand among believe-AI readers ($p = 0.062$). It is rejected for demand among believe-real readers ($p = 0.005$) and for quality among believe-AI readers ($p = 0.023$).

The main source of the rejection for demand is easy to identify. In the easy-AI arm, respondents who did *not* recognise the image as AI ($r = 0$, despite being shown an obviously synthetic image) report newsletter demand of 0.474, well above the 0.391 of believe-real respondents in the control arm. Having concluded that even a low-quality image was a real photograph, these readers appear to update strongly in the outlet's favour. We refer to this as the *hidden-AI bonus*. The model cannot generate it: it assigns every believe-real reader the same demand D^+, regardless of the arm that produced the belief.

Remark B.9 (Implication for welfare analysis). Because the welfare calculations assign every believe-real reader the pooled demand D^+, they understate consumer surplus in the easy-AI arm, and should be read as lower bounds on the gains from undetected AI imagery. The calculations do not depend on the structural parameters π, $\mu^{\pm}$, or b: they use only the pooled $D^{\pm}$ and arm-level attribution rates.

B.14.2 Cross-Equation Proportionality

In the model, demand and quality both respond to the same posterior gap $\mu^+ - \mu^-$, with slopes Δv and b. The ratio of the demand gap to the quality gap must therefore equal $\Delta v/b$ in every arm:

$$\frac{D^+ - D^-}{Q^+ - Q^-} = \frac{\Delta v}{b} = \text{constant}. \tag{35}$$

Table B.7 reports the ratio arm by arm.

Table B.7: Cross-Equation Proportionality Test: $\Delta v/b$ by Arm

Arm	$D^+ - D^-$	$Q^+ - Q^-$	Ratio	Model prediction
C (control, no label)	0.099	0.621	0.159	0.115
LC (control, label)	0.126	0.521	0.242	
HAI (hard AI, no label)	0.074	0.660	0.112	
LHAI (hard AI, label)	−0.062	0.588	−0.105	
EAI (easy AI, no label)	0.137	0.548	0.250	
LEAI (easy AI, label)	0.144	0.340	0.424	
Pooled	0.062	0.536	0.115	0.115

Notes: $D^+ - D^-$ and $Q^+ - Q^-$ are the within-arm gaps in mean demand and quality between believe-real and believe-AI readers. The model prediction is the pooled ratio $(D^+ - D^-)/(Q^+ - Q^-) = \widehat{\Delta v}/\hat{b}$. In the LHAI arm the demand gap is negative, a sign reversal the model does not predict; it is consistent with a label on a hard-to-detect AI image partly restoring the outlet's credibility among readers who recognise the image.

The pooled ratio (0.115) equals the model prediction by construction. Across arms, however, the ratio ranges from −0.105 (LHAI) to 0.424 (LEAI), although no arm-level confidence interval is tight enough to reject equality with the pooled value. The most notable deviation is in LHAI, where labeling a hard-to-detect AI image *reverses* the sign of the demand gap between recognition groups. This is consistent with the transparency bonus described in Section B.7, but incompatible with a strict cross-equation restriction. The model is therefore best understood as a population-average approximation: the arm-level variation in the ratio reflects heterogeneous responses to the interaction of labels and photorealism, which the baseline framework does not model.

B.15 Welfare Analysis

This section uses the belief-conditional demand estimates to evaluate welfare. We derive consumer surplus as a function of the recognition rate (Section B.15.1), compute it under a set of counterfactual scenarios (Sections B.15.2 and B.15.3), and then study the value of detection technology (Section B.15.4), the optimal labeling policy (Section B.15.5), and the interpretation

of the welfare metric (Section B.15.6).

B.15.1 Consumer Surplus Under Uniform Costs

With subscription costs distributed $c \sim \text{Uniform}[0, c_{\max}]$ and $c_{\max}$ normalized to one, a reader with posterior μ subscribes whenever her cost falls below $D(\mu)$, and her expected surplus is

$$\text{CS}(\mu) = \int_0^{D(\mu)} \big(D(\mu) - c\big)\, dc = \frac{D(\mu)^2}{2}. \tag{36}$$

Readers who recognise the image as AI ($r = 1$) demand D^-, and the rest ($r = 0$) demand D^+. If a share $\alpha \equiv \Pr(r = 1)$ of readers recognises the image as AI, average consumer surplus is therefore

$$\text{CS}(\alpha) = \alpha \frac{(D^-)^2}{2} + (1 - \alpha) \frac{(D^+)^2}{2}, \tag{37}$$

where α depends on the image technology and the disclosure regime. Equation (37) requires only D^- and D^+, estimated directly in the heterogeneity analysis, and α, taken from the arm-level attribution rates. It does not depend on π, so the welfare analysis is unaffected by uncertainty about the calibrated prior.

A note on units. The subscription cost c is measured in utility units on $[0, c_{\max}]$. Because the experiment involves no price variation, $c_{\max}$ is not identified. We normalize $c_{\max} = 1$, so consumer surplus is expressed in units of $c_{\max}$ per reader. This has two consequences. First, *percentage* changes in consumer surplus do not depend on $c_{\max}$; these are our headline welfare measures. Second, we express monetary equivalents of surplus changes only as explicit multiples of $c_{\max}$, illustrated for $c_{\max} \in [£5, £50]$, a range that brackets plausible per-reader valuations of a weekly newsletter. Outlet profit, by contrast, is measured directly in pounds, using an externally calibrated subscriber lifetime value $m = £15$, an image-cost saving $\delta = £150$ (Section B.12), and a per-article labeling cost $\kappa_\ell = £5$. We never add consumer surplus and profit at an arbitrary exchange rate. Tables B.8 and B.10 report the two separately, and the policy result in Section B.15.5 holds for *every* $c_{\max} \geq 0$, so no summation is needed.

B.15.2 Baseline Welfare Objects

The baseline is an outlet publishing only real images. Its recognition rate is the false-positive rate in the control arm, $\alpha^{\text{real}} = p_H = 0.367$, and baseline consumer surplus is

$$\text{CS}^{\text{base}} = \text{CS}(p_H) = 0.367 \times \frac{(0.3219)^2}{2} + 0.633 \times \frac{(0.3835)^2}{2} = 0.0656. \tag{38}$$

Under full AI adoption, consumer surplus lies between two extremes:

$$\mathrm{CS}^{\alpha=1} = \frac{(D^-)^2}{2} = 0.0518 \quad \text{(every reader recognises the AI image),} \tag{39}$$

$$\mathrm{CS}^{\alpha=0} = \frac{(D^+)^2}{2} = 0.0735 \quad \text{(no reader recognises it).} \tag{40}$$

B.15.3 Counterfactual Scenarios

Table B.8 reports consumer surplus and outlet profit under seven scenarios. Each scenario is defined by its recognition rate α, taken from the arm-level attribution data or from the detection-capability parameterization of Section B.11.

Table B.8: Counterfactual Welfare by Scenario ($R_{\text{eff}} = 1{,}000$)

Scenario	α	$\bar{D}$	CS (norm.)	ΔCS (%)	Profit (£)
Real images (baseline)	0.367	0.361	0.0656	—	5,263
Full AI, hard, no label	0.441	0.356	0.0639	−2.5	5,195
Full AI, hard, labeled	0.520	0.351	0.0622	−5.1	5,122
Full AI, easy, no label	0.812	0.333	0.0559	−14.8	4,851
Full AI, easy, labeled	0.865	0.330	0.0547	−16.5	4,803
Full AI, always recognised ($\alpha = 1$)	1.000	0.322	0.0518	−21.0	4,678
Full AI, never recognised ($\alpha = 0$)	0.000	0.384	0.0735	+12.2	5,602

Notes: Consumer surplus is computed from equation (37) using the empirical D^-, D^+, and arm-level α, in units of $c_{\max}$ per reader; ΔCS (%) is relative to the real-images baseline and does not depend on $c_{\max}$. Outlet profit $= \bar{D} \times m \times R_{\text{eff}} - \delta$ is in pounds, with $m = £15$, $\delta = £150$, and $R_{\text{eff}} = 1{,}000$. The two columns are in different units and are not summed. Bootstrap 95% CI for the surplus loss under hard AI with no label: $[-4.8\%, -0.5\%]$ (500 resamples).

Three results stand out. First, switching entirely to photorealistic AI images without labels (hard AI, no label) reduces consumer surplus by 2.5% relative to the real-images baseline, with a 95% bootstrap confidence interval of $[-4.8\%, -0.5\%]$. This is the welfare cost of AI adoption at current detection rates. Second, the loss rises steeply with how recognisable the imagery is. Easy-to-detect AI images reduce surplus by 14.8%, nearly six times as much, because far more readers recognise them and move to the low-demand state D^-. Third, if no reader ever recognised the AI images ($\alpha = 0$), measured consumer surplus would *rise* by 12.2%, because every reader would update upward on quality. This is a decision-utility gain generated by a mistaken belief; Section B.15.6 discusses how to interpret it.

B.15.4 Value of Detection Technology

We now ask how welfare changes as readers become better at detecting AI. Let $\lambda \in [0,1]$ index detection capability, as in Section B.11, with $\lambda = 0$ corresponding to current detection and $\lambda = 1$ to perfect detection. Under full adoption of hard photorealistic AI images, the technology an outlet would plausibly choose, the recognition rate is

$$\alpha^{\mathrm{AI}}(\lambda) = p_L + (1 - p_L)\,\lambda = 0.441 + 0.559\,\lambda, \tag{41}$$

so $\alpha^{\mathrm{AI}}(0) = p_L = 0.441$ and $\alpha^{\mathrm{AI}}(1) = 1$. Relative to the real-images baseline, the change in consumer surplus is

$$\Delta\mathrm{CS}(\lambda) = \mathrm{CS}\big(\alpha^{\mathrm{AI}}(\lambda)\big) - \mathrm{CS}(p_H) = \frac{(D^-)^2 - (D^+)^2}{2} \cdot \big(\alpha^{\mathrm{AI}}(\lambda) - p_H\big). \tag{42}$$

Since $D^- < D^+$, the first factor is negative, so once the outlet has adopted AI, consumer surplus falls steadily as detection improves: from -2.5% at $\lambda = 0$ to -21.0% at $\lambda = 1$ (Table B.9). Each unit improvement in detection changes surplus by

$$\frac{\partial\,\Delta\mathrm{CS}}{\partial\lambda} = \frac{(D^-)^2 - (D^+)^2}{2} \cdot (1 - p_L) \approx -0.0121 \tag{43}$$

per reader per article, in units of $c_{\max}$. A ten-percentage-point improvement therefore costs $0.0012\,c_{\max}$ per reader, or $£1.21 \times c_{\max}$ per article per 1,000 readers (bootstrap 95% CI: $[-1.79, -0.59] \times c_{\max}$). This amounts to £6 at $c_{\max} = £5$ and £61 at $c_{\max} = £50$. Over the same improvement, the outlet loses $0.062 \times 0.559 \times 0.1 \times m \times 1{,}000 \approx £52$ in revenue per article per 1,000 readers at $m = £15$, a figure that does not depend on $c_{\max}$. Better detection thus acts as a modest but persistent tax on AI-adopting outlets, through both reader surplus and revenue.

B.15.5 Optimal Disclosure Policy

Let $\gamma \in [0,1]$ denote the label intensity chosen by a social planner, from $\gamma = 0$ (no labeling) to $\gamma = 1$ (mandatory labeling of all AI images). Labels raise recognition, and we take the effect to be linear in γ:

$$\alpha^{\mathrm{AI}}(\gamma) = p_L + \gamma\Delta p_{\mathrm{AI}}, \qquad \Delta p_{\mathrm{AI}} = p_L^{\mathrm{label}} - p_L = 0.520 - 0.441 = 0.079. \tag{44}$$

Total welfare is

$$W(\gamma; R_{\mathrm{eff}}, c_{\max}) = R_{\mathrm{eff}}\Big[c_{\max} \cdot \mathrm{CS}\big(\alpha^{\mathrm{AI}}(\gamma)\big) + m \cdot D\big(\alpha^{\mathrm{AI}}(\gamma)\big)\Big] - \delta - \gamma\kappa_\ell, \tag{45}$$

where multiplying by $c_{\max}$ converts consumer surplus into pounds.

Table B.9: Consumer Surplus as a Function of Detection Capability λ

λ	α^{AI}	$\bar{D}$	CS (norm.)	ΔCS (%)
0.0	0.441	0.356	0.0639	−2.5
0.1	0.497	0.353	0.0627	−4.3
0.2	0.553	0.349	0.0615	−6.2
0.3	0.609	0.346	0.0603	−8.0
0.4	0.665	0.343	0.0591	−9.9
0.5	0.721	0.339	0.0579	−11.7
0.6	0.776	0.336	0.0567	−13.6
0.7	0.832	0.332	0.0554	−15.4
0.8	0.888	0.329	0.0542	−17.3
0.9	0.944	0.325	0.0530	−19.1
1.0	1.000	0.322	0.0518	−21.0

Notes: $\alpha^{\mathrm{AI}}(\lambda) = 0.441 + 0.559\lambda$ from equation (41); $\lambda = 0$ is current detection. CS from equation (37). ΔCS relative to the real-images baseline ($\mathrm{CS}^{\mathrm{base}} = 0.0656$).

Proposition B.10 (Optimal label intensity: robust to the welfare weighting)**.** *Suppose $D^- < D^+$, so that $\partial \mathrm{CS}/\partial\alpha = [(D^-)^2 - (D^+)^2]/2 < 0$ and $\partial D/\partial\alpha = D^- - D^+ < 0$. Then for every $c_{\max} \geq 0$, $m \geq 0$ (not both zero), $\kappa_\ell \geq 0$, and $R_{\mathrm{eff}} > 0$, $W(\gamma)$ is strictly decreasing in γ and the planner's optimum is $\gamma^* = 0$ (no mandatory labeling). The same conclusion holds for any planner objective $\omega_1 \mathrm{CS} + \omega_2 \Pi$ with non-negative weights.*

Proof. Differentiating equation (45):

$$\frac{\partial W}{\partial \gamma} = R_{\mathrm{eff}} \left[c_{\max} \frac{\partial \mathrm{CS}}{\partial \alpha} + m \frac{\partial D}{\partial \alpha} \right] \Delta p_{\mathrm{AI}} - \kappa_\ell .$$

Both partial derivatives are strictly negative, so the bracket is strictly negative for any non-negative $(c_{\max}, m)$ not both zero; with $\kappa_\ell \geq 0$, $\partial W/\partial\gamma < 0$ and $\gamma^* = 0$. Because consumer surplus and profit each fall in α, the sign of $\partial W/\partial\gamma$ does not depend on how the two components are weighted or monetised. □

The proposition shows that, in the quality-constant world of the experiment, mandatory labeling reduces welfare regardless of how consumer surplus is valued: the unidentified $c_{\max}$ plays no role in the sign. The reason is that a label moves readers toward the low-demand belief state D^- without telling them anything about the newsletter's value, so it reduces reader surplus and outlet revenue at the same time. For concreteness, at $R_{\mathrm{eff}} = 1{,}000$ moving from $\gamma = 0$ to $\gamma = 1$ costs about £73 in subscription revenue, $1.72 \times c_{\max}$ in consumer surplus, and $\kappa_\ell = £5$ in operational cost per article. Table B.10 reports the components separately by outlet scale.

Table B.10: Welfare Cost of Mandatory Labeling by Outlet Scale (components reported separately)

R_{eff}	γ^*	$\Delta\Pi(\gamma{:}\,0\to1)$ (£)	$\Delta\text{CS}(\gamma{:}\,0\to1)$ ($\times c_{\max}$)	κ_ℓ (£)	Policy
500	0.000	−36	−0.86	5	No labeling
1,000	0.000	−73	−1.72	5	No labeling
2,000	0.000	−146	−3.43	5	No labeling
5,000	0.000	−365	−8.58	5	No labeling
10,000	0.000	−730	−17.2	5	No labeling

Notes: $\Delta\Pi = -\Delta p_{\text{AI}}\,(D^+ - D^-)\,m R_{\text{eff}}$ with $\Delta p_{\text{AI}} = 0.079$, $D^+ - D^- = 0.062$, and $m = £15$. $\Delta\text{CS} = \Delta p_{\text{AI}}\,[(D^-)^2 - (D^+)^2]/2 \cdot R_{\text{eff}}$, in units of $c_{\max}$; multiply by $c_{\max}$ to obtain pounds. Since both components are negative for all $c_{\max} \geq 0$, $\gamma^* = 0$ at every scale and for every monetisation of consumer surplus (Proposition B.10).

This result depends on the experiment's quality-constant design. If AI imagery were instead associated with lower editorial standards, as the model assumes in Section B.5, the planner's objective would include the cost of misled readers, and mandatory labeling could be optimal, as in Goldberg and Lam (2025).

The unambiguous component: the false-positive penalty on authentic photographs. The welfare cost of AI-origin labels depends on one's view of belief-based utility (Section B.15.6). One component of the current regime, however, is a loss under *any* metric: readers penalising authentic photographs that they mistake for synthetic ones. For an outlet publishing only real images, the false-positive rate is $p_H = 0.367$. If authentication technology eliminated false positives ($\alpha = 0$), consumer surplus would rise by

$$\text{Cost of false positives} = \text{CS}(0) - \text{CS}(p_H) = 0.0735 - 0.0656 = 0.0080 \tag{46}$$

per reader in units of $c_{\max}$, or 12.2% of baseline consumer surplus, and the outlet would gain $p_H\,(D^+ - D^-)\,m \approx £0.34$ in revenue per reader. This is the same 12.2% as in the never-recognised AI scenario of Table B.8, because in both cases every reader ends up believing the image is real. What differs is whether that belief is correct. For authentic photographs it is, so eliminating false positives moves beliefs toward the truth *and* raises surplus, and the gain counts under any welfare metric. For AI images the belief is mistaken, so the analogous "cost of noise" cannot be signed without a stance on belief-based utility: false negatives raise decision-utility surplus (the hidden-AI bonus) while moving beliefs away from the truth. We therefore report only the false-positive component, which is the quantity relevant for image-authentication standards such as C2PA Content Credentials. Their welfare value is unambiguous, whereas the value of AI-origin labels depends on whether AI use is associated with quality.

B.15.6 Interpreting the Welfare Metric: Decision Utility and Equilibrium Consistency

All welfare objects in this appendix are *decision-utility* measures: they value a subscription at the reader's believed willingness to pay, $\bar{c}(\mu)$. Under this metric, the $+12.2\%$ surplus when no reader recognises the AI images (Table B.8) is generated by a mistaken belief. The harm from labeling, in turn, consists of moving readers to a pessimistic belief that is not payoff-relevant in our quality-constant design. Labeling here does not redistribute surplus to anyone; it destroys decision-utility surplus outright.

Two further issues deserve explicit treatment.

Consistency with rational expectations. The model's information structure sits uneasily with the experimental design. Assumption B.2 ($p_L > p_H$) makes a reader's downgrade upon recognition *rational* only if, among the outlets she faces, AI use is genuinely associated with lower editorial type. Our experiment deliberately removes that association: image content, composition, and quality are held constant, so image origin carries no information about quality by construction. Readers' inferences may therefore be sound as beliefs about the market yet wrong about the particular outlet in front of them. In a full rational-expectations equilibrium with no link between AI use and quality, the belief-conditional demand gap could not persist: readers would learn that recognising AI tells them nothing about value, and D^- would converge to D^+. This leaves two readings. Under reading (i), readers hold market-level beliefs that our design makes locally wrong. The decision-utility losses we compute are then transitional and would fade as audiences learn that content-equivalent AI imagery is harmless. Under reading (ii), the recognition penalty reflects a direct, categorical distaste for AI provenance, a preference rather than an inference. Decision utility is then exactly the right metric, and the losses are permanent. Three pieces of evidence favour reading (ii): the implausibly large quality sensitivity $\hat{b} = 7.1$ (Section B.13.4), the insensitivity of the conditional penalty to how diagnostic the signal is, and the in-sample violation of diagnosticity in Experiment 1.

Robustness of the disclosure result. Within the quality-constant world, the optimal-disclosure conclusion does not depend on whether welfare is measured by decision or experienced utility. Under experienced utility with quality held constant, the newsletter's true value is the same in every belief state, so beliefs about image origin are payoff-irrelevant. A label that shifts those beliefs conveys no useful information, yet still distorts subscription decisions away from the full-information benchmark and mechanically reduces outlet revenue. Under decision utility, the label destroys surplus directly (Proposition B.10). On either metric, then, $\gamma^* = 0$ in the quality-constant environment. The two metrics diverge only in how they value *unrecognised* AI (the $+12.2\%$ scenario): a gain under decision utility, and roughly neutral under experienced utility. Both agree that mandatory labeling becomes valuable precisely when AI use is associated with lower quality, which reconciles our findings with Goldberg and Lam (2025), as discussed in the Introduction. Throughout, we flag which objects are robust to the choice of metric: the false-

positive cost of Section B.15.5 is, while the size of the detection-technology tax (Section B.15.4) is a decision-utility object.

B.15.7 Bootstrap Confidence Intervals

Table B.11 reports bootstrap 95% confidence intervals for the key welfare objects, based on 500 resamples with replacement of the full Experiment 2 dataset.

Table B.11: Bootstrap 95% Confidence Intervals for Welfare Objects

Welfare object	Estimate	95% CI	
		Lower	Upper
CS baseline (real images)	0.0655	0.0595	0.0715
CS under full AI, always recognised	0.0517	0.0455	0.0577
CS under full AI, never recognised	0.0735	0.0658	0.0827
CS under hard AI, no label	0.0639	0.0582	0.0696
CS loss from hard AI, no label (%)	−2.46%	−4.78%	−0.54%
Demand gap $D^{+} - D^{-}$	0.0617	0.0313	0.0883
Value of 10pp detection gain (per 1,000 readers, $\times\, c_{\max}$)	−1.21	−1.79	−0.59

Notes: Percentile bootstrap intervals, 500 resamples. CS objects are in normalized units ($c_{\max} = 1$); their monetary equivalents are multiples of $c_{\max}$. Profit objects use $m = £15$. The CS loss from hard AI adoption is statistically distinguishable from zero at the 5% level.

C Appendix Tables

Table C.1: Label Effects on Beliefs and Behaviour: Benchmarking Across Domains

Panel A AI & Media Labels				
This paper (Exps. 1 & 2) UK, Prolific; $N = 7{,}066$	AI-image label on news article	Newsletter sign-up; perceived outlet quality	Avg. null: +0.03 pp (s.e. 0.011); recognised AI: −10 pp vs. unrecognised; 2SLS quality penalty: −0.35 SD	Aggregate null on behaviour; ±0.45–0.55 SD on quality conditional on recognition group; belief-driven mechanism confirmed by photorealism IV
Altay and Gilardi 2024 *PNAS Nexus; US+UK; N* = 4,976	"AI-generated" headline label	Perceived accuracy; sharing intention	Accuracy: ≈ −0.10 SD ($\frac{1}{3}\times$ "false" label)	Holds for true and human-made headlines; driven by overestimation of AI automation
Wittenberg et al. 2025 *PNAS Nexus; US; N* = 7,579	Process-based vs. harm-based AI-image label	Perceived accuracy; engagement	Accuracy: −0.13 to −0.25 SD; engagement: near-null	Harm-based labels more effective on accuracy; neither type clearly reduces sharing
Toff and Simon 2024 *J. Comm.; US; N* ≈ 2,000	AI-generated news disclosure	Perceived trustworthiness	−0.12 to −0.18 SD	Lower trust even when accuracy not rated as worse
Li and Yang 2024 *JMIR; China; N* = 800	AIGC label on health posts	Accuracy; credibility; sharing	Main effect null (*ns*)	Content category and veracity interactions dominate
Panel B Misinformation & Fact-check Labels				
Pennycook et al. 2020 *Mgmt. Sci.; US; N* ≈ 3,000	"Disputed" tag on fake-news headline	Perceived accuracy (labelled & unlabelled)	Tagged false: −0.20 SD; implied truth (unlabelled): +0.05 SD	Unlabelled false headlines rated more accurate when some content is tagged ("implied truth effect")
Pennycook and Rand 2021 *Nature; US; N* ≈ 5,000	Accuracy-nudge prompt	Sharing discernment (true−false)	$d \approx 0.14$	Driven by lift in true-news sharing, not suppression of false sharing
Martel and Rand 2023 *Curr. Op. Psych.; meta-review*	Warning / disputed labels	Accuracy perception; sharing	$d \approx 0.10$–0.25	Larger for politically congruent content; broadly effective across party lines
Panel C Nutrition & Food Labels				
Muller et al. 2021 *J. Acad. Mkt. Sci.; 1.9M transactions*	Nutri-Score, traffic light, GDA	Purchase of top-third healthy items	Lab: +30–40%; field: +14%; 17× smaller in field	Key benchmark for lab-to-field attenuation; habit and inattention dominate real-world settings
Bleich et al. 2017 *Obesity; systematic review*	Calorie / menu labels in restaurants	Calories purchased	−20 to −78 kcal/meal in lab; null in most natural experiments	Highly variable; controlled effects rarely replicate in large field settings
Hainmueller et al. 2015 *Rev. Econ. Stat.;* ≈ 200,000 *tx*	Fair Trade certification (coffee)	Unit sales; price premium	Sales +9–10%; 8% WTP premium	Real-purchase field experiment; aspirational label with positive demand effect
Bastounis et al. 2021 *Nutrients; meta-analysis, 47 DCEs*	Environmental sustainability label	Willingness to pay (WTP)	WTP premium +8–25%	Stated-preference context; real-purchase WTP substantially lower
Panel D Tobacco & Health Warning Labels				
Brewer et al. 2021 *JAMA Netw. Open; RCT; N* = 2,149	Graphic warning label (GWL) on cigarette pack	Quit cognitions; cessation	Quit cognitions: +0.10–0.15 SD; consumption: null	GWLs shift intentions but not actual smoking behaviour at 3 months
Hammond 2011 *Tob. Control; systematic review*	Pictorial vs. text-only tobacco warning	Risk perception; quit intention	$d \approx 0.20$–0.35	Consistent awareness effects; behavioural cessation effects smaller and mixed
Panel E Origin & Provenance Labels				
Sampalean et al. 2023 *Food Policy; systematic review, 20+ experiments*	Country-of-origin (COO) / PDO-PGI	Choice probability; quality evaluation	Domestic preference: large (conjoint); actual purchase lift: small–moderate	Large stated-preference effects; substantial behavioural gap in field settings

Notes: Effect sizes in Cohen's *d* or standard deviations (SD) where available; otherwise percentage-point (pp) or percentage changes. tx = transactions; GDA = Guideline Daily Amount; DCE = Discrete Choice Experiment; WTP = Willingness to Pay. The aggregate null on newsletter sign-up mirrors the dominant pattern in Panels B–D. Quality-perception effects (−0.35 SD via IV; ±0.45–0.55 SD conditional on recognition) are at the high end of Panel B and comparable to Panel D.

Table C.2: Summary Statistics: Experiment 1-2 vs. UK population

Variable (definition)	**Experiment 1**	**Experiment 2**	**UK Population**
Gender (share female)	52.3%	51.5%	51.1%
Age (years)	46.7	44.4	40.7 (median)
Ethnicity (share White)	88.3%	87.3%	83.0%
UK-born (share born in the UK)	89.1%	93.5.3%	84.0%
Region (share living in the South East)	14.8%	16.1%	15.6%
Income (share earning > £40K)	41.7%	59.2%	25.0%
Employment (share employed)	75.8%	76.1%	75.1%
Education (share with university degree)	62.3%	64.0%	33.8%
Party (share previously voting Labour)	36.6%	27.1%	33.7%

Notes: *Experiment 1* and *Experiment 2* figures refer to our survey respondents in the two waves, respectively. *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 C.3: Balance across Treatments: Group Means, Global and Pairwise Tests (Experiment 1)

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 C.4: Manipulation Check (Experiment 1)

	Manipulation ($AI = Yes$)			
	(1)	(2)	(3)	(4)
	OLS	OLS	Logit	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 C.5: AI and Label Effects on Requesting the Newsletter (Experiment 1)

	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 C.6: AI and Label Effects on Requesting the Newsletter, by Recognition (Experiment 1)

	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 C.7: Mechanisms: Effects of AI and Label on Perceived Outlet Quality (Experiment 1)

	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 C.8: AI and Label Effects on Requesting the Newsletter and on Perceived Outlet Quality (**IV estimates, Experiment 1**)

	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 C.9: Mechanisms (Alternative): Effects of AI and Label on Perceived Bias, Complexity, and Engagement (Experiment 1)

	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 C.10: 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 C.11: 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).

Table C.12: Balance across Treatments: Group Means, Global and Pairwise Tests (Experiment 2)

Variable	Control	Control + Label	Easy AI	Easy AI + Label	Hard AI	Hard AI + Label	Global_p	Min_pairwise_p
Gender (share of 'Female')	0.54	0.51	0.52	0.51	0.51	0.5	0.52	1.00
Age	44.93	44.73	43.63	44.14	44.7	44.08	0.57	1.00
Ethnicity (share of 'White')	0.86	0.88	0.85	0.89	0.88	0.89	0.93	1.00
UKborn (share of 'Yes')	0.93	0.93	0.94	0.94	0.92	0.96	0.05	0.05
Region (share of 'South East')	0.17	0.15	0.16	0.14	0.16	0.18	0.98	1.00
Income (share of '£20,000 - £29,999')	0.17	0.16	0.15	0.16	0.17	0.15	0.64	1.00
Employment (share of 'Employed full time')	0.55	0.6	0.58	0.56	0.58	0.57	0.57	0.69
Education (share of 'University or higher')	0.62	0.66	0.65	0.66	0.62	0.64	0.49	1.00
Political (share of 'Liberal')	0.38	0.41	0.4	0.37	0.35	0.42	0.24	0.19
Party (share of 'Labour')	0.28	0.25	0.31	0.27	0.23	0.29	0.16	0.02
InterestNews (share of 'Interested')	0.43	0.41	0.4	0.39	0.37	0.38	0.60	0.42
ConsumptionNews (share of 'Every day')	0.55	0.57	0.53	0.56	0.54	0.54	0.94	1.00
Newsletter (share of 'Yes')	0.16	0.17	0.15	0.15	0.14	0.14	0.47	0.80
Post-treatment variables								
ArticleTime (seconds reading article)	80.01	76.05	80.25	75.49	80.05	73.69	0.78	1.00
Manipulation (share believe AI)	0.37	0.36	0.81	0.87	0.44	0.52	0.00	0.00
OriginCorrect (share identify AI/Not)	0.63	0.64	0.81	0.87	0.44	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. Categorical rows show the share of the modal category. The column *Min pairwise p* shows the smallest Holm-adjusted p-value from pairwise t- or proportion tests between any two treatment arms.

Table C.13: AI and Label Effects on Manipulation and Correct Recognition (Experiment 2)

	Manipulation (*AI = Yes*)				**OriginCorrect** (*= Yes*)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Label Control	-0.002	0.0010			0.007	0.006		
	(0.026)	(0.025)			(0.026)	(0.026)		
Easy AI	0.446***	0.451***	0.446***	0.451***	0.177***	0.174***	0.177***	0.174***
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)
Easy AI + Label	0.496***	0.499***			0.232***	0.231***		
	(0.022)	(0.022)			(0.022)	(0.022)		
Hard AI	0.073***	0.072***	0.073***	0.072***	-0.194***	-0.192***	-0.194***	-0.192***
	(0.026)	(0.025)	(0.026)	(0.025)	(0.026)	(0.026)	(0.026)	(0.026)
Hard AI + Label	0.154***	0.156***			-0.111***	-0.113***		
	(0.026)	(0.026)			(0.026)	(0.026)		
Label			-0.002	0.0010			0.007	0.006
			(0.026)	(0.025)			(0.026)	(0.026)
Easy AI × Label			0.051	0.048			0.048	0.051
			(0.032)	(0.032)			(0.032)	(0.032)
Hard AI × Label			0.083**	0.084**			0.076**	0.073**
			(0.037)	(0.036)			(0.037)	(0.037)
Pseudo R^2	0.141	0.152	0.141	0.152	0.084	0.091	0.084	0.091
Observations	4,196	4,196	4,196	4,196	4,196	4,196	4,196	4,196
Image FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Mean DV	0.561	0.561	0.561	0.561	0.652	0.652	0.652	0.652

Notes: *Manipulation* equals 1 if respondents identified the image as AI-generated, 0 otherwise. *OriginCorrect* equals 1 if respondents correctly recognized whether the image was AI-generated or not, 0 otherwise. *Easy AI* equals 1 for treatment arms Easy AI Unlabeled (EAI) and Easy AI Labeled (LEAI), 0 otherwise; *Hard AI* equals 1 for Hard AI Unlabeled (HAI) and Hard AI Labeled (LHAI), 0 otherwise; *Label* equals 1 if a label was present (Label Control, LEAI, LHAI), 0 otherwise. Controls include age, gender, education, income, political orientation, newsletter subscription, and AI familiarity. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table C.14: AI and Label Effects on Manipulation and Correct Recognition (Logit, Experiment 2)

	Manipulation ($AI = Yes$)				**OriginCorrect** ($= Yes$)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Label Control	-0.005	0.008			0.033	0.028		
	(0.113)	(0.114)			(0.113)	(0.114)		
Easy AI	2.05***	2.11***	2.05***	2.11***	0.919***	0.914***	0.919***	0.914***
	(0.127)	(0.130)	(0.127)	(0.130)	(0.125)	(0.126)	(0.125)	(0.126)
Easy AI + Label	2.44***	2.49***			1.32***	1.33***		
	(0.138)	(0.140)			(0.137)	(0.137)		
Hard AI	0.310***	0.308***	0.310***	0.308***	-0.803***	-0.803***	-0.803***	-0.803***
	(0.110)	(0.111)	(0.110)	(0.111)	(0.111)	(0.112)	(0.111)	(0.112)
Hard AI + Label	0.644***	0.666***			-0.459***	-0.475***		
	(0.110)	(0.111)			(0.110)	(0.111)		
Label			-0.005	0.008			0.033	0.028
			(0.113)	(0.114)			(0.113)	(0.114)
Easy AI × Label			0.388**	0.371**			0.373**	0.389**
			(0.187)	(0.189)			(0.185)	(0.187)
Hard AI × Label			0.339**	0.350**			0.310**	0.300*
			(0.156)	(0.158)			(0.156)	(0.158)
Pseudo R^2	0.146	0.157	0.146	0.157	0.088	0.096	0.088	0.096
Observations	4,196	4,196	4,196	4,196	4,196	4,196	4,196	4,196
Image FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Mean DV	0.561	0.561	0.561	0.561	0.652	0.652	0.652	0.652

Notes: Each column reports logit estimates. *Manipulation* equals 1 if respondents identified the image as AI-generated, 0 otherwise. *OriginCorrect* equals 1 if respondents correctly recognized whether the image was AI-generated or not, 0 otherwise. *Easy AI* equals 1 for treatment arms Easy AI Unlabeled (EAI) and Easy AI Labeled (LEAI), 0 otherwise; *Hard AI* equals 1 for Hard AI Unlabeled (HAI) and Hard AI Labeled (LHAI), 0 otherwise; *Label* equals 1 if a label was present (Label Control, LEAI, LHAI), 0 otherwise. Controls include age, gender, education, income, political orientation, newsletter subscription, and AI familiarity. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table C.15: AI and Label Effects on Requesting the Newsletter (Experiment 2)

	Demand (*Newsletter* = *Yes*)											
	(1) OLS	(2) Logit	(3) OLS	(4) Logit	(5) OLS	(6) Logit	(7) OLS	(8) Logit	(9) OLS	(10) Logit	(11) OLS	(12) Logit
Label Control	-0.014	-0.061	-0.020	-0.093	-0.004	-0.019	-0.010	-0.045	-0.032	-0.159	-0.038	-0.202
	(0.025)	(0.112)	(0.025)	(0.115)	(0.033)	(0.138)	(0.032)	(0.140)	(0.040)	(0.198)	(0.039)	(0.203)
Easy AI	0.018	0.076	0.019	0.088	0.096*	0.395**	0.104**	0.450**	0.195***	0.835***	0.215***	0.967***
	(0.026)	(0.111)	(0.025)	(0.115)	(0.050)	(0.201)	(0.049)	(0.210)	(0.052)	(0.222)	(0.051)	(0.231)
Easy AI + Label	-0.024	-0.106	-0.025	-0.120	0.066	0.270	0.048	0.204	0.164***	0.709***	0.159***	0.721***
	(0.025)	(0.113)	(0.025)	(0.116)	(0.057)	(0.231)	(0.057)	(0.244)	(0.059)	(0.249)	(0.059)	(0.263)
Hard AI	-0.006	-0.028	-0.006	-0.031	-0.009	-0.040	-0.018	-0.085	0.089**	0.400**	0.092**	0.432**
	(0.026)	(0.112)	(0.025)	(0.115)	(0.034)	(0.143)	(0.033)	(0.147)	(0.038)	(0.173)	(0.037)	(0.178)
Hard AI + Label	-0.021	-0.093	-0.016	-0.074	-0.090***	-0.398***	-0.086***	-0.405**	0.009	0.041	0.025	0.112
	(0.025)	(0.113)	(0.025)	(0.115)	(0.034)	(0.154)	(0.034)	(0.158)	(0.038)	(0.182)	(0.037)	(0.187)
Manipulation					-0.098***	-0.439***	-0.111***	-0.517***				
					(0.037)	(0.168)	(0.036)	(0.173)				
Label Control × Manipulation					-0.027	-0.140	-0.028	-0.158				
					(0.051)	(0.242)	(0.050)	(0.247)				
Easy AI × Manipulation					-0.043	-0.149	-0.042	-0.158				
					(0.060)	(0.259)	(0.060)	(0.270)				
Easy AI + Label × Manipulation					-0.047	-0.181	-0.021	-0.076				
					(0.066)	(0.282)	(0.066)	(0.296)				
Hard AI × Manipulation					0.023	0.106	0.046	0.214				
					(0.052)	(0.234)	(0.051)	(0.241)				
Hard AI + Label × Manipulation					0.161***	0.723***	0.169***	0.794***				
					(0.051)	(0.233)	(0.050)	(0.239)				
OriginCorrect									0.098***	0.439***	0.111***	0.517***
									(0.037)	(0.168)	(0.036)	(0.173)
Label Control × OriginCorrect									0.027	0.140	0.028	0.158
									(0.051)	(0.242)	(0.050)	(0.247)
Easy AI × OriginCorrect									-0.240***	-1.03***	-0.264***	-1.19***
									(0.060)	(0.258)	(0.060)	(0.268)
Easy AI + Label × OriginCorrect									-0.244***	-1.06***	-0.243***	-1.11***
									(0.066)	(0.282)	(0.066)	(0.296)
Hard AI × OriginCorrect									-0.174***	-0.773***	-0.176***	-0.820***
									(0.052)	(0.235)	(0.051)	(0.243)
Hard AI + Label × OriginCorrect									-0.035	-0.156	-0.053	-0.240
									(0.051)	(0.234)	(0.050)	(0.240)
Pseudo R^2	0.002	0.002	0.038	0.038	0.010	0.010	0.046	0.047	0.010	0.010	0.046	0.047
Observations	4,196	4,196	4,196	4,196	4,196	4,196	4,196	4,196	4,196	4,196	4,196	4,196
Image FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	No	No	Yes	Yes	No	No	Yes	Yes
Mean DV	0.349	0.349	0.349	0.349	0.349	0.349	0.349	0.349	0.349	0.349	0.349	0.349

Notes: *Demand* equals 1 if respondents said they wish to receive the Newsletter, 0 otherwise. *Manipulation* equals 1 if respondents identified the image as AI-generated, 0 otherwise. *OriginCorrect* equals 1 if respondents correctly recognized whether the image was AI-generated or not, 0 otherwise. Control is the omitted treatment category in all columns. Even-numbered columns report logit estimates; odd-numbered columns report OLS estimates. Controls include age, gender, education, income, political orientation, newsletter subscription, and AI familiarity. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table C.16: Mechanisms: Effects of AI and Label on Perceived Outlet Quality (Experiment 2)

	z-index	**Not Accurate**	**Not Trustworthy**	**Low Quality**	**Biased**	**Very Complex**	**Not Entertaining**	**No Researcher Trust**
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Label Control	0.057	-0.007	-0.011	0.009	0.013	0.052*	-0.081*	0.015
	(0.076)	(0.037)	(0.038)	(0.037)	(0.043)	(0.030)	(0.043)	(0.032)
Easy AI	0.681***	-0.172***	-0.170***	-0.158***	-0.045	0.024	-0.180***	-0.124***
	(0.094)	(0.035)	(0.037)	(0.035)	(0.051)	(0.036)	(0.051)	(0.028)
Easy AI + Label	0.499***	-0.202***	-0.214***	-0.130***	-0.084	0.033	-0.246***	-0.142***
	(0.082)	(0.032)	(0.032)	(0.040)	(0.057)	(0.041)	(0.058)	(0.025)
Hard AI	0.622***	-0.188***	-0.182***	-0.159***	-0.071*	0.024	-0.085**	-0.114***
	(0.069)	(0.029)	(0.030)	(0.029)	(0.038)	(0.026)	(0.038)	(0.025)
Hard AI + Label	0.659***	-0.202***	-0.205***	-0.163***	-0.087**	0.053*	-0.125***	-0.135***
	(0.070)	(0.029)	(0.029)	(0.030)	(0.040)	(0.028)	(0.041)	(0.024)
OriginCorrect	0.649***	-0.189***	-0.191***	-0.164***	-0.083**	0.023	-0.170***	-0.122***
	(0.067)	(0.028)	(0.029)	(0.029)	(0.038)	(0.025)	(0.038)	(0.024)
Label Control × OriginCorrect	-0.107	0.009	0.010	0.009	0.025	-0.033	0.092*	-0.004
	(0.093)	(0.040)	(0.041)	(0.041)	(0.054)	(0.038)	(0.054)	(0.035)
Easy AI × OriginCorrect	-1.20***	0.314***	0.320***	0.343***	0.089	-0.006	0.224***	0.223***
	(0.108)	(0.040)	(0.042)	(0.041)	(0.060)	(0.042)	(0.060)	(0.033)
Easy AI + Label × OriginCorrect	-0.957***	0.353***	0.351***	0.267***	0.124*	-0.042	0.264***	0.240***
	(0.098)	(0.038)	(0.038)	(0.044)	(0.064)	(0.046)	(0.065)	(0.030)
Hard AI × OriginCorrect	-1.29***	0.392***	0.402***	0.357***	0.213***	-0.041	0.165***	0.273***
	(0.095)	(0.040)	(0.041)	(0.041)	(0.053)	(0.036)	(0.053)	(0.035)
Hard AI + Label × OriginCorrect	-1.22***	0.376***	0.365***	0.350***	0.156***	-0.052	0.218***	0.237***
	(0.093)	(0.038)	(0.039)	(0.040)	(0.053)	(0.037)	(0.053)	(0.032)
R^2	0.115	0.074	0.073	0.077	0.064	0.040	0.055	0.046
Observations	4,196	4,196	4,196	4,196	4,196	4,196	4,196	4,196
Image FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	0	0.146	0.156	0.171	0.422	0.144	0.517	0.1

Notes: Each column reports OLS estimates of treatment effects on perceived outlet quality measures. *z-index* is a standardized composite index averaging *NotAccurate*, *NotTrustworthy*, and *LowQuality* (higher = worse perceived quality). *NotAccurate*, *NotTrustworthy*, and *LowQuality* are binary indicators equal to 1 if the respondent selected one of the two lowest categories on the respective 1–5 Likert scale, 0 otherwise. *Biased*, *VeryComplex*, *NotEntertaining*, and *NoResearcherTrust* are similarly coded binary indicators. *OriginCorrect* equals 1 if respondents correctly recognized whether the image was AI-generated or not, 0 otherwise. Control is the omitted treatment category. Controls include age, gender, education, income, political orientation, newsletter subscription, and AI familiarity. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table C.17: AI and Label Effects on Requesting the Newsletter and Perceived Outlet Quality **(IV estimates, Experiment 2)**

	Demand (*Newsletter* = *Yes*)				**z-index**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Manipulation	0.007	0.010			-0.366***	-0.351***		
	(0.036)	(0.036)			(0.066)	(0.066)		
OriginCorrect			0.015	0.009			-0.348***	-0.345***
			(0.050)	(0.049)			(0.091)	(0.091)
Observations	4,196	4,196	4,196	4,196	4,196	4,196	4,196	4,196
F-test (1st stage), Manipulation	171.1	174.1			171.1	174.1		
F-test (1st stage), OriginCorrect			91.2	90.3			91.2	90.3
Wald (2nd stage)	0.033	0.084	0.089	0.037	30.5	28.6	14.8	14.5
Image FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Mean DV	0.349	0.349	0.349	0.349	0	0	0	0

Notes: Each column reports 2SLS estimates where *Manipulation* or *OriginCorrect* is instrumented by treatment assignment. *Demand* equals 1 if respondents said they wish to receive the Newsletter, 0 otherwise. *z-index* is a standardized composite index of perceived outlet quality (higher = better). *Manipulation* equals 1 if respondents identified the image as AI-generated, 0 otherwise. *OriginCorrect* equals 1 if respondents correctly recognized whether the image was AI-generated or not, 0 otherwise. Even-numbered columns include controls for age, gender, education, income, political orientation, newsletter subscription, and AI familiarity. All columns include image fixed effects. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table C.18: Mechanism: AI and Label Effects on Perceived Outlet Quality (**IV estimates, Experiment 2**)

	z-index	**Not Accurate**	**Not Trustworthy**	**Low Quality**	**Biased**	**Very Complex**	**Not Entertaining**	**No Researcher Trust**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
OriginCorrect	-0.345***	0.108***	0.084**	0.126***	-0.046	0.0005	-0.166***	0.053
	(0.091)	(0.038)	(0.039)	(0.039)	(0.050)	(0.036)	(0.052)	(0.032)
Observations	4,196	4,196	4,196	4,196	4,196	4,196	4,196	4,196
F-test (1st stage), OriginCorrect	90.3	90.3	90.3	90.3	90.3	90.3	90.3	90.3
Wald (2nd stage)	14.5	8.20	4.74	10.1	0.836	0.0002	10.2	2.68
Image FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean DV	0	0.146	0.156	0.171	0.422	0.144	0.517	0.1

Notes: Each column reports 2SLS estimates where *OriginCorrect* is instrumented by treatment assignment (Easy AI, Hard AI, Label, and their interactions). *z-index* is a standardized composite index of perceived outlet quality (higher = better). *NotAccurate*, *NotTrustworthy*, and *LowQuality* are binary indicators equal to 1 if the respondent selected one of the two lowest categories on the respective 1–5 Likert scale, 0 otherwise. *Biased*, *VeryComplex*, *NotEntertaining*, and *NoResearcherTrust* are similarly coded binary indicators. *OriginCorrect* equals 1 if respondents correctly recognized whether the image was AI-generated or not, 0 otherwise. Controls include age, gender, education, income, political orientation, newsletter subscription, and AI familiarity. All columns include image fixed effects. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Robust standard errors in parentheses.

Table C.19: Journalists' Sample Characteristics

Characteristic	Category	N	Value
Age	Mean (SD)	132	35.1 (10.4)
Gender	Female	79	59.8%
Gender	Male	50	37.9%
Gender	Non-binary	2	1.5%
Gender	Prefer not to say	1	0.8%
Country (top 10)	United States of America	40	35.1%
Country (top 10)	United Kingdom	24	21.1%
Country (top 10)	South Africa	12	10.5%
Country (top 10)	Germany	8	7%
Country (top 10)	Mexico	8	7%
Country (top 10)	Spain	7	6.1%
Country (top 10)	Canada	4	3.5%
Country (top 10)	Italy	4	3.5%
Country (top 10)	Poland	4	3.5%
Country (top 10)	France	3	2.6%
Employment status	Working full time now	94	71.2%
Employment status	Working part time now	27	20.5%
Employment status	Student	6	4.5%
Employment status	Unemployed	3	2.3%
Employment status	Temporarily laid off	1	0.8%
Employment status	Other	1	0.8%
Currently a journalist	Yes	101	76.5%
Currently a journalist	No, but I have in the past	26	19.7%
Currently a journalist	No, I have never worked as journalist	5	3.8%
Outlet type	Online Newspaper	38	29.2%
Outlet type	Freelance	34	26.2%
Outlet type	TV	20	15.4%
Outlet type	Daily Newspaper	16	12.3%
Outlet type	Weekly Newspaper	14	10.8%
Outlet type	Blog	8	6.2%
Tenure in journalism	More than 5 years	60	83.3%
Tenure in journalism	Less than 1 year	12	16.7%
Role	Editor	54	41.5%
Role	Reporter	51	39.2%
Role	Opinion Writer	18	13.8%
Role	Columnist	7	5.4%

Table C.20: Journalists' Reported Reasons for Using AI in Their Work

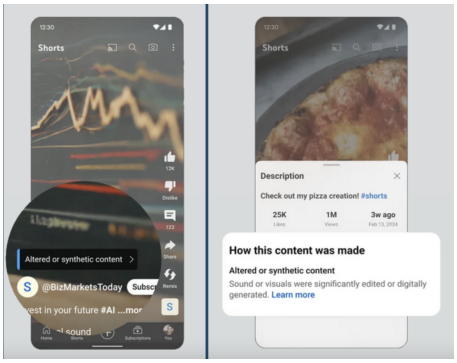
Topic Category	% of Respondents	Illustrative Example
Efficiency and Time Pressure	35.7%	*"Journalists are using AI to reduce the time consumption it takes to research, gather and present their work as they have tight deadlines to meet."*
Research, Data, and Information Processing	31%	*"AI can be used as a tool by subeditors to generate quick alt copy for images (for visually impaired readers) or by reporters to collate data."*
Professional Concerns and Ethical Issues	29.4%	*"Any journalist who uses AI in their articles is a disgrace to their profession. It's lazy, insulting and morally wrong."*
Workflow Facilitation and Content Scaling	27%	*"AI has come to facilitate journalistic work, and in an environment where a very high workload is constantly demanded, AI is a tool that allows journalists to meet their bosses' demands."*
Image and Content Generation	22.2%	*"Journalists use it to generate pictures which are sometimes hard to find and AI can assist with creating those images."*
Writing and Editing Support	22.2%	*"I think journalists use AI to make their articles more intriguing. They make their articles to seem more in depth and interesting."*
Transparency, Training, and Ethical Frameworks	12.7%	*"I think journalists are using AI in their articles as a way to speed up the writing process. When working to tight deadlines, using AI can be a way to ensure accuracy in spelling and grammar, which will result in less time spent proof reading."*

Notes: Categories are not mutually exclusive; respondents could mention multiple themes. Percentages reflect the share of the 126 respondents whose answer contained at least one reference to the given theme.

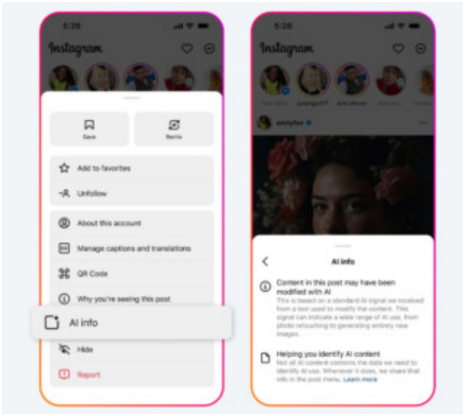
D Appendix Figures

Figure D.1: Disclosure of AI-generated content on major platforms

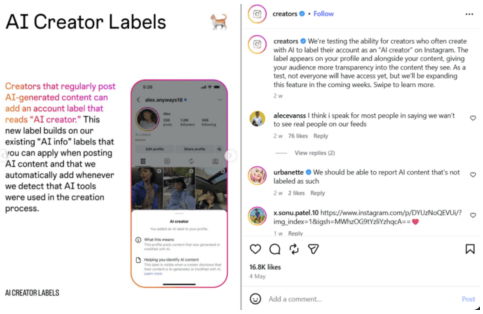
(a) Viewer-facing label on YouTube



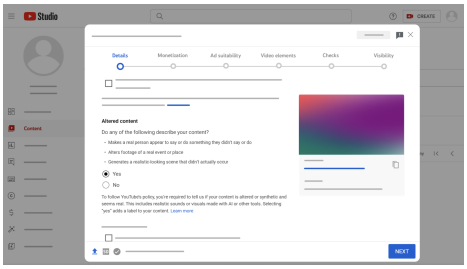
(b) Post-level label on Instagram



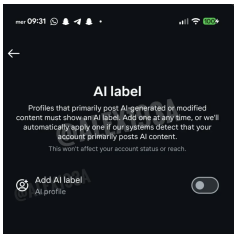
(c) Account-level label on Instagram



(d) Creator disclosure flow on YouTube



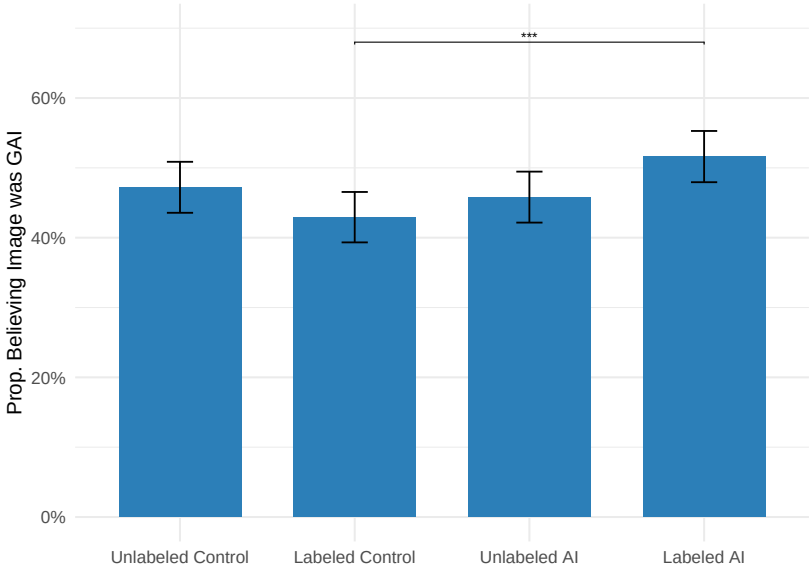
(e) Creator disclosure setting on Instagram



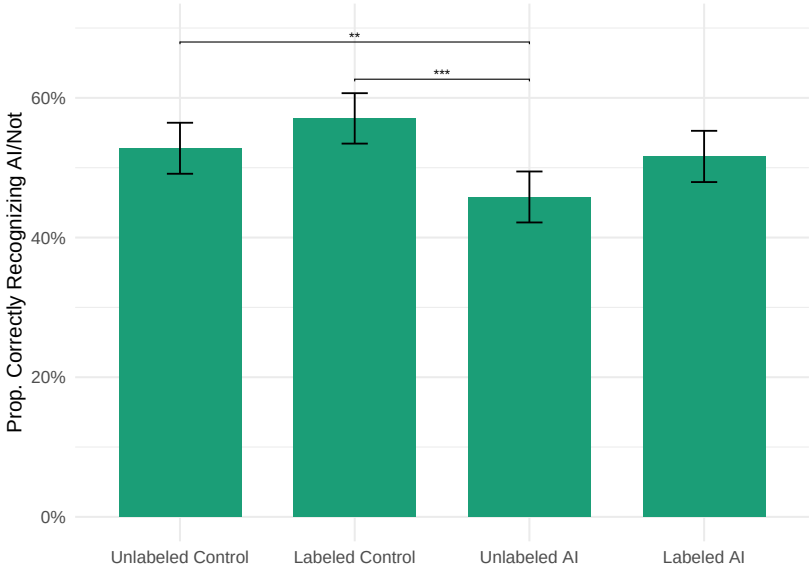
Notes: Panels (a)–(c) show labels as they appear to viewers; Panels (d)–(e) show the corresponding disclosure interfaces presented to the content creator. Panel (a): YouTube displays an "Altered or synthetic content" label on the video player for sensitive topics, and reports in the expanded description that sound or visuals were significantly edited or digitally generated. Panel (b): Instagram places an "AI info" entry in the post overflow menu, which opens a panel reporting that content in the post may have been modified with AI. Panel (c): Instagram additionally offers an account-level "AI creator" label, displayed on the profile and alongside the account's content. Panel (d): YouTube requires creators to answer whether their content makes a real person appear to say or do something they did not, alters footage of a real event or place, or generates a realistic-looking scene that did not occur; selecting "Yes" attaches the viewer-facing label shown in Panel (a). Panel (e): Meta requires profiles that primarily post AI-generated or modified content to carry an AI label, applied either by the account holder or automatically. Screenshots retrieved from platform documentation and newsrooms. All marks remain the property of their respective owners and are reproduced here for scholarly commentary.

Figure D.2: Manipulation Check Results (Experiment 1)

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

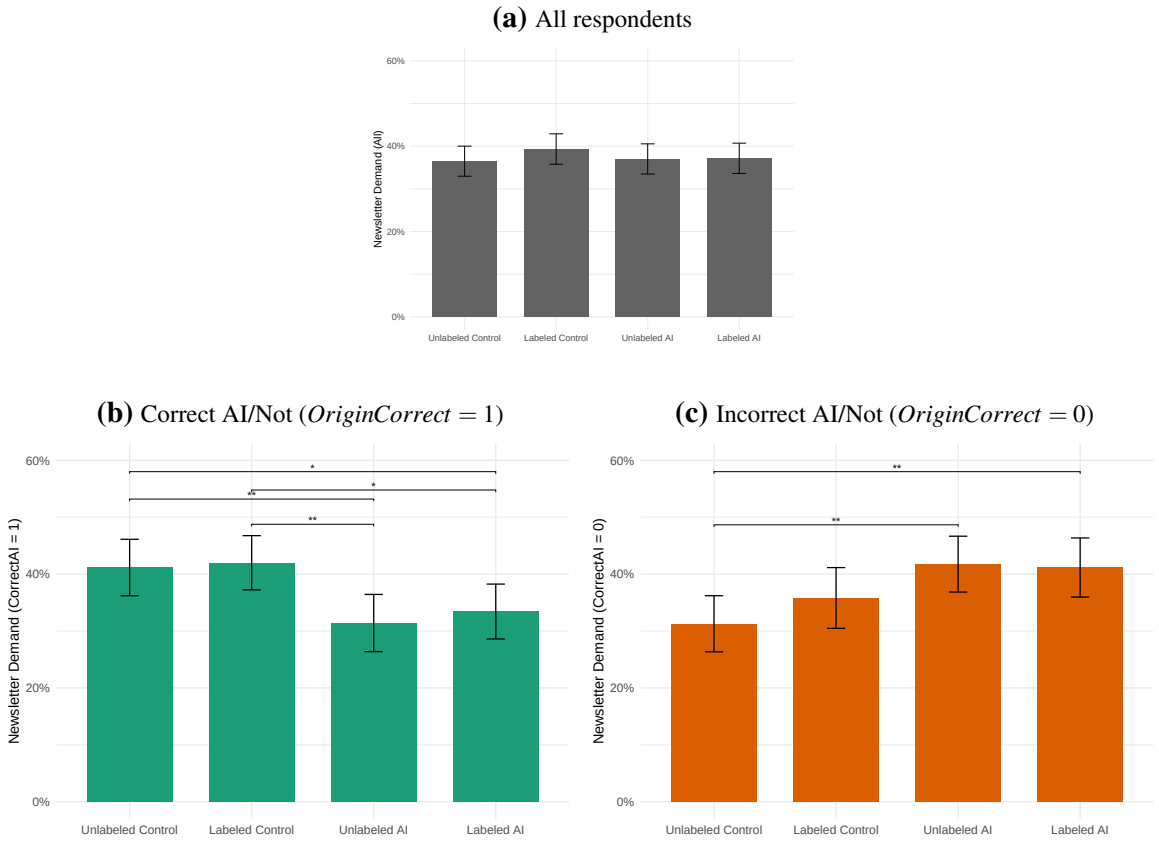


(b) Share Identifying AI/Non-AI Correctly (*OriginCorrect = 1*) by Treatment Arm



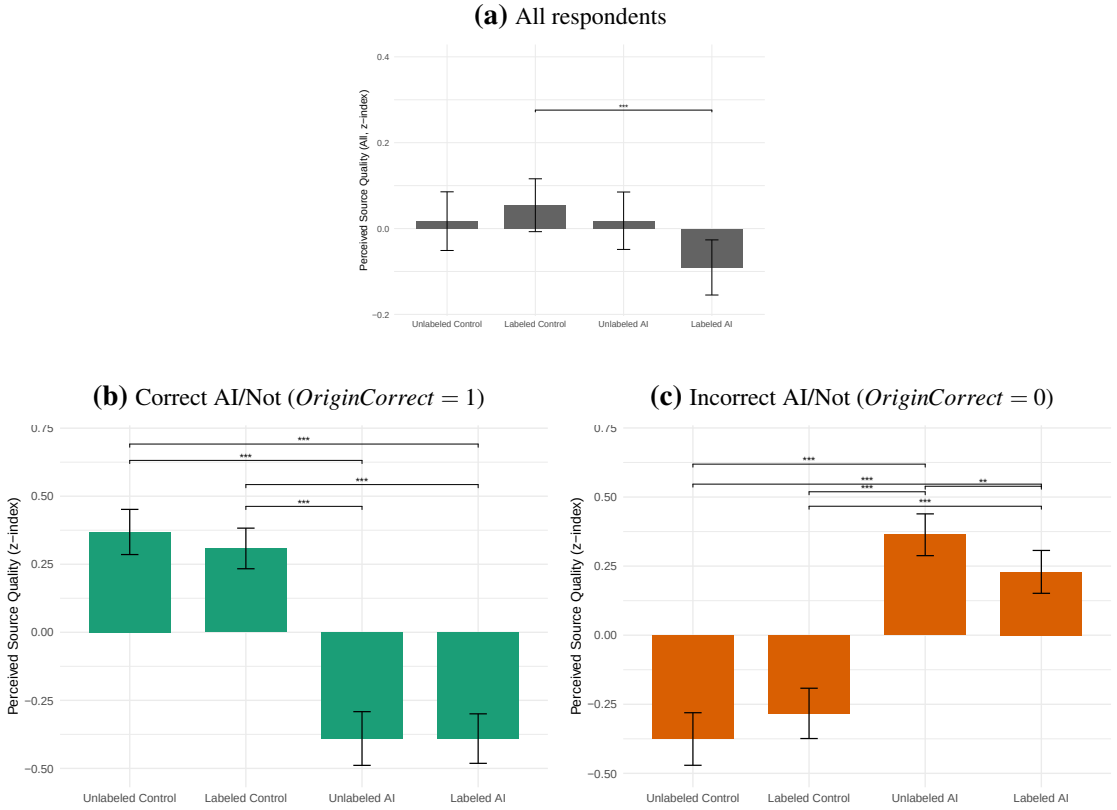
Notes: Panel (a) shows the share of respondents reporting that the image was AI-generated (*Manipulation = 1*); Panel (b) shows the share who correctly recognized whether the image was AI-generated or not (*OriginCorrect = 1*). Error bars show 95% confidence intervals; stars above brackets denote significance levels (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$). Only significant pairwise comparisons are reported. Experiment 2 results appear in Figure 5.

Figure D.3: Newsletter Demand by Treatment and AI Recognition Accuracy (Experiment 1)



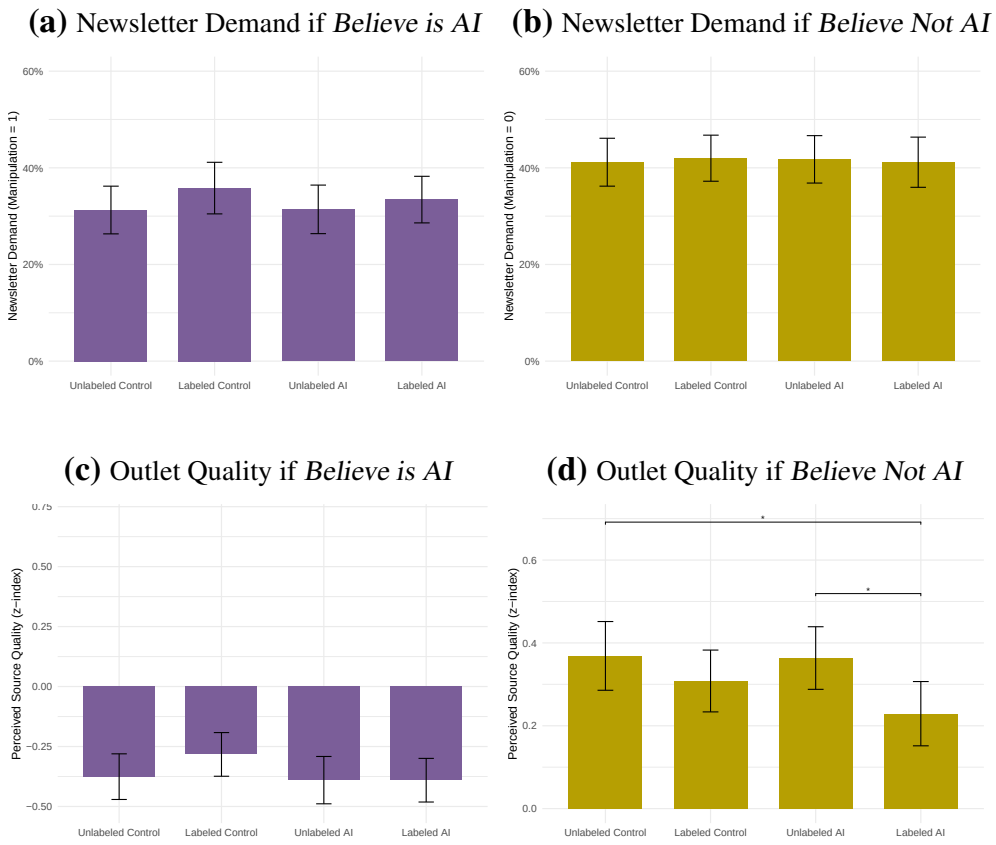
Notes: Each panel shows the share of respondents demanding the newsletter by treatment arm (Experiment 1, $N = 2{,}870$). Panel (a) includes all respondents; Panels (b) and (c) split the sample by whether respondents correctly identified whether the image was AI-generated ($OriginCorrect = 1$) or failed to do so ($OriginCorrect = 0$). Error bars represent 95% confidence intervals; stars above brackets denote significance levels (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$). Only significant pairwise comparisons are reported. Experiment 2 results appear in Figure 6.

Figure D.4: Perceived Outlet Quality by Treatment and AI Recognition Accuracy (Experiment 1)



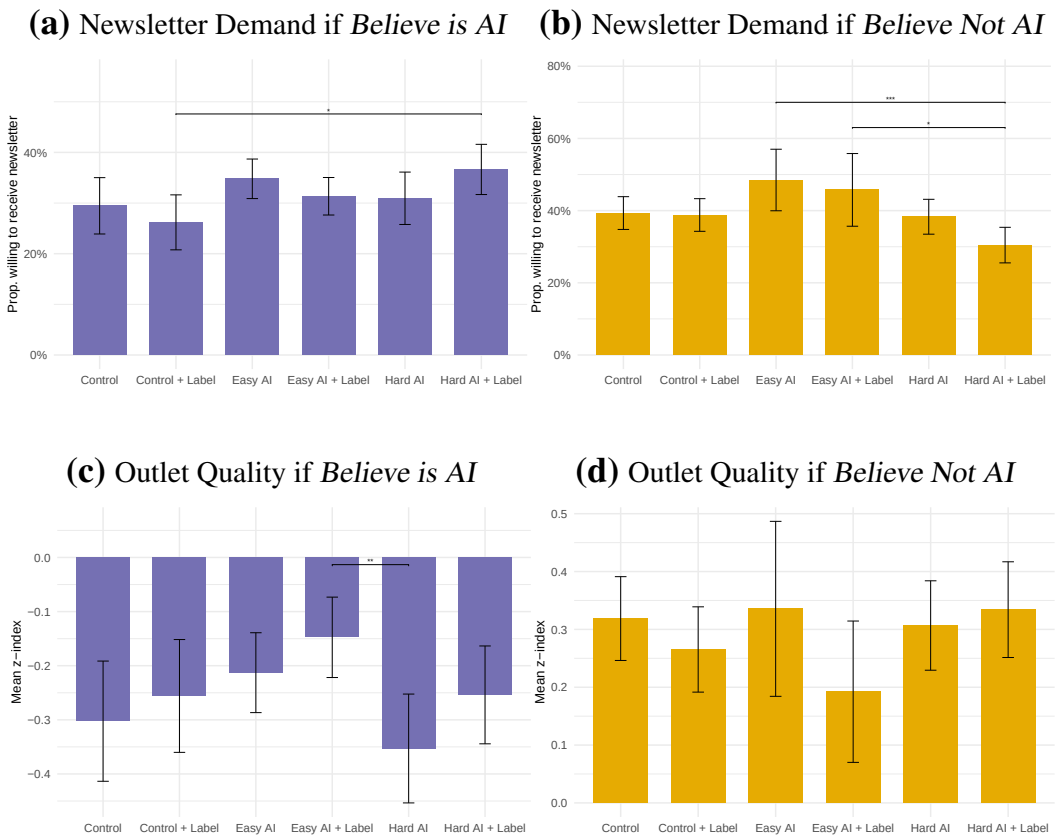
Notes: Each panel shows the average perceived outlet quality by treatment arm (Experiment 1, $N = 2{,}870$). Perceived outlet quality is a composite index averaging standardized responses to three 1–5 scale items: *Accuracy*, *Trustworthiness*, and *Quality*. Panel (a) includes all respondents; Panels (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 (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$). Only significant pairwise comparisons are reported. Experiment 2 results appear in Figure 7.

Figure D.5: Newsletter Demand and Perceived Outlet Quality by Treatment and AI Beliefs (Experiment 1)



Notes: Panels (a) and (b) show the share of respondents demanding the newsletter by treatment arm. Panels (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) restrict to respondents who believed the image was AI-generated ($Manipulation = 1$); panels (b) and (d) restrict to those who did not ($Manipulation = 0$). Error bars represent 95% confidence intervals; stars above brackets denote significance levels (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$).

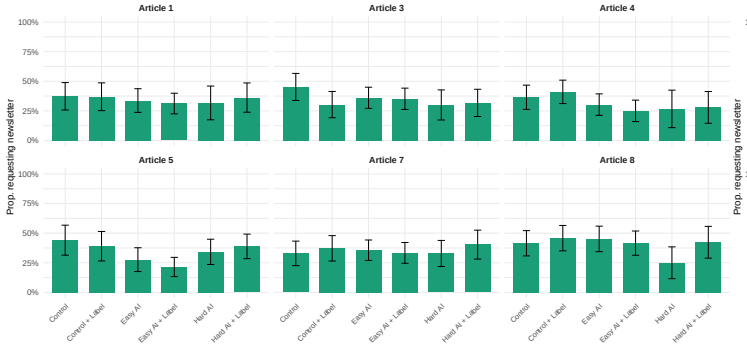
Figure D.6: Newsletter Demand and Perceived Outlet Quality by Treatment and AI Beliefs (Experiment 2)



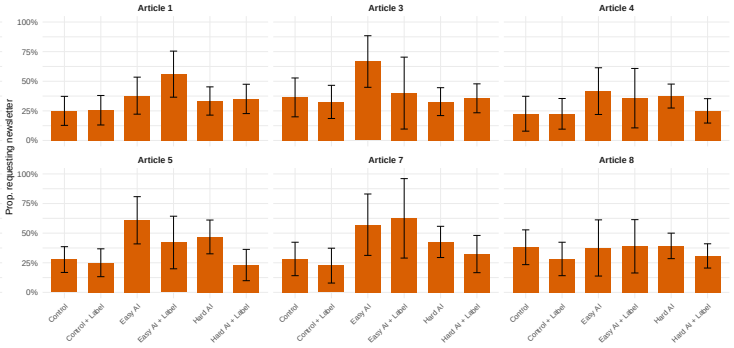
Notes: Panels (a) and (b) show the share of respondents demanding the newsletter by treatment arm. Panels (c) and (d) show the average perceived outlet quality by treatment arm. Perceived outlet quality is a composite index averaging standardized responses to three 1–5 scale items: *Accuracy*, *Trustworthiness*,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

Figure D.7: Newsletter Demand by Treatment and Article (Experiment 2)

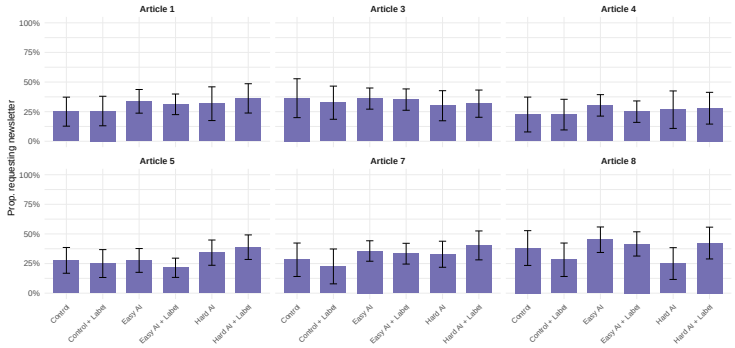
(a) Newsletter Demand if *Correct AI/Not*



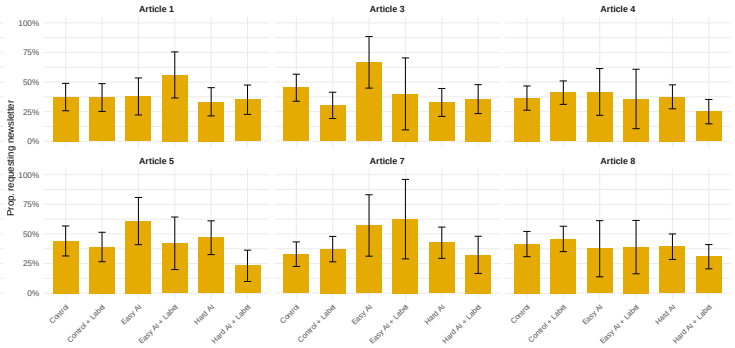
(b) Newsletter Demand if *Incorrect AI/Not*



(c) Newsletter Demand if *Believe is AI*



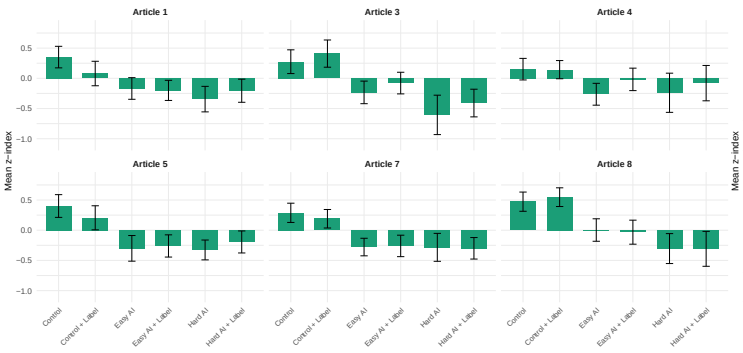
(d) Newsletter Demand if *Believe Not AI*



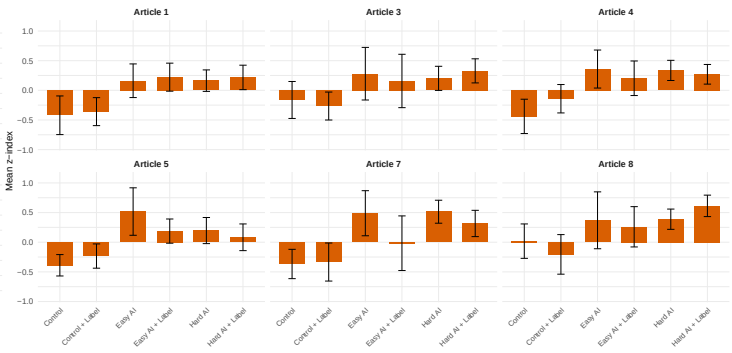
Notes: Each panel corresponds to one of the six articles shown to respondents. Panels (a)–(b) and (c)–(d) split the sample by whether respondents correctly identified the origin of the image (*OriginCorrect* $= 1$ and $= 0$ respectively) or by whether they believed it was AI-generated (*Manipulation* $= 1$ and $= 0$ respectively). Bar heights show the share of respondents requesting the newsletter within each treatment arm and article. Error bars represent 95% confidence intervals. Pairwise significance tests are not reported as the within-article sample size is insufficient to power them reliably.

Figure D.8: Perceived Outlet Quality by Treatment and Article (Experiment 2)

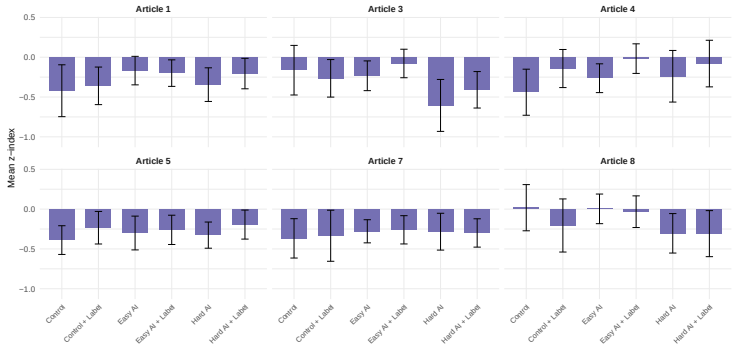
(a) Outlet Quality if *Correct AI/Not*



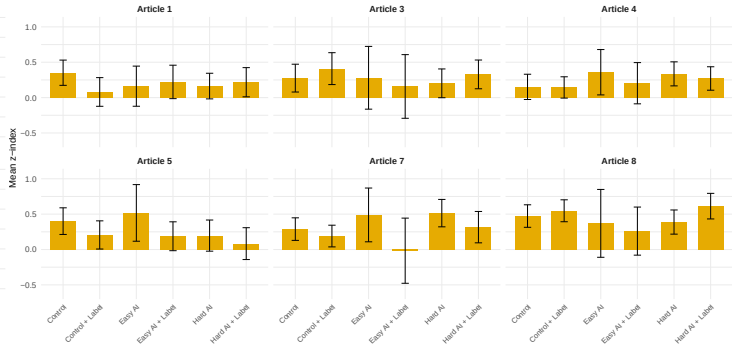
(b) Outlet Quality if *Incorrect AI/Not*



(c) Outlet Quality if *Believe is AI*



(d) Outlet Quality if *Believe Not AI*



Notes: Each panel corresponds to one of the six articles shown to respondents. Panels (a)–(b) and (c)–(d) split the sample by whether respondents correctly identified the origin of the image (*OriginCorrect* $= 1$ and $= 0$ respectively) or by whether they believed it was AI-generated (*Manipulation* $= 1$ and $= 0$ respectively). Bar heights show mean z-index within each treatment arm and article. Error bars represent 95% confidence intervals. Pairwise significance tests are not reported as the within-article sample size is insufficient to power them reliably.

E Image Generation Prompts

Experiment 1. The AI image was generated using DALL·E. The following prompt was used to produce a photorealistic version closely matching the original wire-service photograph:

> "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."

Experiment 2. Image generation followed a two-step protocol designed to separate scene description from image generation. In the first step, a dedicated instance of ChatGPT was provided with the original article page and instructed to produce a detailed, literal description of the accompanying photograph covering objects, setting, composition, perspective, lighting, and salient visual elements without any reference to provenance, realism, or artistic style.

In the second step, a separate instance of ChatGPT Image (with no access to the original photograph) was provided with the textual description from step one and instructed to generate an image matching the described scene and composition. At this stage, photorealism was varied across two versions: one prompted to be highly photorealistic (*very realistic AI*), closely resembling a professional news photograph, and one prompted to be intentionally less photorealistic (*not very realistic AI*), with visual cues making AI generation more apparent. The depicted scene, composition, and semantic content were held constant across versions.

F Instructions

A full print preview of the Qualtrics survey for Experiment 1 (including all instructions and blocks) is available at the following link.

The instructions of Experiment 2 are nearly identical to those of Experiment 1, but with two additional treatments. The complete set of instructions and article pictures can be found in the following link.

Finally, the complete set of instructions for the journalists' survey can be found in the following link.