

The Credibility Cost of Mandatory AI Disclosure

Experimental Evidence from Online News in Seven Countries

Ivan Muzic^{a,b} Paulo Sankovic^b

^aFaculty of Economics and Business, University of Rijeka ^bNewsLabs Group

October 2026

The EU AI Act requires publishers to label news text generated by AI, unless a person has thoroughly reviewed it and holds editorial responsibility for it. Publishers fear that such labels cost them credibility. We test this fear in a randomized experiment with 679 readers in five EU and two non-EU countries. All respondents read the same local news story on a realistic mobile news page, and only the byline changed: a journalist, a journalist “with” an AI assistant, or “created by AI”. These versions match the cases in the Act: no label, a voluntary label for reviewed text, and a mandatory label for AI-generated text. The AI labels did not lower how credible, professional or well made readers found the article, and they did not reduce the share of readers who would read it to the end. We can rule out penalties larger than about 0.2 points on a 7-point scale. Many readers assumed that AI was involved anyway: even when only a journalist was named, 62% believed that AI had helped produce the article. Many also did not notice the label: two thirds could not recall how the author was shown. Still, the label changed what readers believed about how the article was made, but not how they rated it: 53% of those who saw “created by AI” thought the article was made mostly or entirely by AI, against 26% of those who saw only a journalist’s name, yet both groups gave it similar ratings of credibility and professionalism. Readers who believed that AI wrote the article rated it lower, but this correlation would predict a penalty three times larger than the one the label caused. For routine news, the label required by the AI Act comes at no measurable cost to credibility.

JEL codes: C90, D83, L51, L82, O33.

Keywords: generative AI, mandatory disclosure, information regulation, news credibility, EU AI Act, survey experiment, instrumental variables, equivalence testing.

1. Introduction

Generative AI has moved into the newsroom. Publishers now use large language models to draft, summarize, translate and adapt routine stories, usually with a journalist who checks and signs the final text (Simon, 2024). Evidence from other writing and service tasks suggests that such tools raise output per worker substantially (Noy and Zhang, 2023; Brynjolfsson et al., 2025). At the same time, regulators ask for transparency. Article 50(4) of the EU Artificial Intelligence Act, which applies from 2 August 2026, requires deployers of an AI system that generates text “published with the purpose of informing the public on matters of public interest” to disclose that the text was artificially generated, unless the content has undergone human review or editorial control and a natural or legal person holds editorial responsibility for it (European Parliament and Council of the European Union, 2024; European Commission, 2026b).

The rule creates an economic choice for publishers. A newsroom can keep a journalist responsible for every AI-assisted text and stay inside the exemption, or it can automate further and label the output. Which choice is cheaper depends on what the label costs. News outlets sell credibility: readers return to sources they trust, and outlets invest in their reputation to protect it (Gentzkow and Shapiro, 2006). If an AI label lowers credibility, mandatory disclosure works like a tax on automated production, and the exemption becomes a reason to keep people in the loop even where they add little. If the label costs little, the rule informs readers at a low price. This paper measures that price.

There are reasons to expect such a cost. Surveys suggest that readers are uneasy about AI in news. Across 28 markets in 2024, only 36 percent of respondents said they were comfortable with news produced mostly by a journalist with some help from AI, and only 19 percent with news produced mostly by AI with some human oversight (Newman et al., 2024). In 2025, Croatia had the lowest comfort with mostly AI-produced news of all 48 markets surveyed, at 9 percent (Newman et al., 2025). Experiments that attach AI labels to headlines find that labels reduce perceived accuracy (Longoni et al., 2022; Altay and Gilardi, 2024) and can lower trust in the outlet (Toff and Simon, 2025). If these effects carry over to real reading, a disclosure rule would impose a credibility cost on newsrooms that use AI, and it could push them either to use AI less or to disclose it less.

We estimate the credibility cost of AI disclosure with a randomized experiment in seven countries: five EU member states that are bound by the AI Act (Croatia, France, Germany, Italy and Spain) and two countries outside the EU that serve as a comparison (the United Kingdom and the United States). In total, 679 adults read the same short local news story about a change in the city bus timetable. The story appeared on a mobile news page that we built for the study and adapted to a real city in each country. The only element we varied was the byline. In arm A, the byline showed a journalist’s name. In arm B, it showed the same name followed by “with NewsLabs AI”, a disclosure of AI assistance. In arm C, there was no journalist; the byline said “This article was created by AI (NewsLabs)”. Arm B corresponds to a voluntary disclosure for content that a journalist reviewed, which the Act does not require. Arm C corresponds to the case in which disclosure is mandatory.

We report three main findings. First, the AI labels did not reduce how readers judged the article. Perceived credibility, quality and professionalism were essentially the same in all three versions. On a standardized index of the three ratings, the estimated effect is -0.050 standard deviations (SD) for arm B

and -0.054 SD for arm C. The 95 percent confidence intervals exclude penalties larger than about 0.24 SD, and when we pool the two labels, which do not differ from each other, an equivalence test rejects effects as large as 0.2 SD in either direction ($p = 0.030$). Readers were also equally willing to read the article: 38 percent in arm A, 38 percent in arm B and 39 percent in arm C would read it to the end, and the labels did not increase the share who would leave the site. These results are robust to randomization inference, alternative estimators and samples, and they are similar across countries and groups of readers. Second, readers assume that AI is involved anyway. Even when the byline showed only a journalist's name, 62 percent believed that AI had helped produce the article, and only 18 percent thought that a human wrote it without AI. This prior is common in every country and in every group of readers we can observe, including those who never use AI at work. Third, most readers did not know which label they had seen. Asked from memory how the author was shown, 66 percent said they did not remember. Only a third of those who saw "created by AI" recalled it, only 17 percent of those who saw "with NewsLabs AI" recalled a tool next to the journalist's name, and no one opened the short explanation of the AI assistant.

Several further results support this picture. The labels did shift stated beliefs about production for some readers: the share who thought the article was produced mostly or entirely by AI was 26 percent under the journalist byline and 53 percent under the "created by AI" label. But this shift did not carry over to ratings. In the cross section, respondents who believed that the article was written entirely by AI rated it about 0.8 SD lower than those who believed that a human wrote it without AI; if that gradient were causal, the belief shift caused by the "created by AI" label would have lowered ratings by about 0.15 SD, three times the actual effect. Using the random labels as instruments for beliefs, we find that this gradient overstates the causal effect of believing that AI wrote an article, most likely because readers who find a text weak infer that a machine wrote it. General attitudes also differ strongly across readers: 72 percent of respondents who never use AI at work prefer news written without AI, against 41 percent of daily users. Yet the reaction to the label does not differ between these groups.

We interpret these findings with a simple model in which readers combine what the byline tells them about how an article was produced with what they learn from reading the text itself. When readers can see the full article, the text is an informative signal of quality, and a label about the production process receives little weight. Two further features of our data limit the scope for a label to matter. Most readers do not process the byline closely, and most already assume that AI is involved.

The paper contributes to three literatures. The first is the economics of disclosure. A large literature studies when firms disclose quality voluntarily and what mandatory disclosure achieves (Grossman, 1981; Milgrom, 1981; Jin and Leslie, 2003; Dranove and Jin, 2010). AI labels disclose a production process rather than a quality level, much like labels for genetically engineered food, where mandatory disclosure did not reduce, and may have improved, consumer attitudes (Kolodinsky and Lusk, 2018). Behavioral work stresses that disclosures work only if people notice and process them (Loewenstein et al., 2014). We provide a direct estimate of the private cost of a disclosure mandate for an information good, with treatment arms that match the cases defined in the law. We also show how limited the processing of a byline label is, and that the label nevertheless shifted beliefs without moving evaluations.

Second, we contribute to work on how audiences respond to AI disclosure in news and other content.

Experiments with headlines and short messages often find penalties (Longoni et al., 2022; Altay and Gilardi, 2024; Toff and Simon, 2025; Wittenberg et al., 2025), while studies that show longer texts often find small or no differences in perceived credibility (Graefe et al., 2018; Henestrosa et al., 2023; Gilardi et al., 2026), and labels do not reduce the persuasive effect of AI-generated messages (Gallegos et al., 2026). Work on bylines suggests that perceived AI contribution, more than the wording of the label, predicts credibility (Jia et al., 2024). We add a multi-country randomized test with full articles in a realistic layout, a direct comparison between an “assisted” and a “created by AI” label, and equivalence tests that quantify how small the effect is. To our knowledge, this is also the first study to use randomly assigned AI labels as instruments for readers’ beliefs about AI authorship. This separates the effect of believing that a text was produced by AI from the reverse inference that drives the cross-sectional correlation reported in earlier work.

Third, we relate to research on algorithm aversion and appreciation (Dietvorst et al., 2015; Logg et al., 2019). Aversion to algorithms is stronger for subjective tasks than for objective ones (Castelo et al., 2019), and recent evidence finds persistent penalties for AI-written creative work (Raj et al., 2026) and emotional marketing messages (Kirk and Givi, 2025). Routine factual news sits at the objective end of this range, and our results suggest that, for such content, aversion to AI authorship has little effect on how a well-made article is judged.

The study was not pre-registered. To limit discretion in the analysis, we fix one primary outcome, report every outcome we collected, correct for multiple testing, and base inference on the known randomization procedure. The anonymized data, the analysis code and the complete experimental materials, including the source code of the survey application and the stimuli in all six languages, are publicly available, so that the design can be replicated independently.

The rest of the paper is organized as follows. Section 2 describes the regulatory background and a simple conceptual framework. Section 3 presents the experimental design, and Section 4 the data and the checks on the validity of the randomization. Section 5 sets out the empirical strategy. Section 6 reports the results, Section 7 discusses their interpretation and limits, and Section 8 concludes.

2. Background and conceptual framework

2.1. AI in newsrooms and the disclosure rule

Article 50 of the AI Act sets transparency obligations for providers and deployers of certain AI systems. Paragraph 4 covers text: deployers of an AI system that generates or manipulates text published to inform the public on matters of public interest must disclose that the text has been artificially generated or manipulated. The obligation does not apply where the AI-generated content “has undergone a process of human review or editorial control” and where a natural or legal person “holds editorial responsibility for the publication of the content” (European Parliament and Council of the European Union, 2024). The disclosure must be made in a clear and distinguishable manner at the latest at the time of first exposure (Article 50(5)). The obligations apply from 2 August 2026; the 2026 amendments to the Act left Article 50(4) and its date of application unchanged (European Parliament and Council of the European Union, 2026).

Two documents clarify how the rule applies to news. The Commission’s Code of Practice on marking and labelling of AI-generated content, published in June 2026, sets out how deployers can label AI-generated text (European Commission, 2026a). The Commission’s guidelines on Article 50, approved in July 2026, specify that the two conditions of the exception are cumulative, that human review means a substantive review of the content by competent people (with fact checking as a minimum), and that a cursory approval does not qualify (European Commission, 2026b). In practice, an article drafted with AI and then reviewed and signed by a journalist is exempt, so disclosure is voluntary. An article generated by AI and published without such review must be labelled.

Our arms map onto this distinction. Arm B (“with NewsLabs AI”) is the kind of disclosure a newsroom might choose for an AI-assisted article that a journalist reviewed. Arm C (“This article was created by AI”) is the kind of label that the rule requires for unreviewed AI text. The UK and the US are not subject to the AI Act and serve as comparison markets.

2.2. A simple model of how a byline label affects evaluations

To fix ideas, consider a reader i who evaluates the quality q of an article. Let $s \in \{0, 1, 2, 3\}$ denote the extent of AI involvement in production, from “written by a human without AI” ($s = 0$) to “entirely AI” ($s = 3$). The reader’s prior is that quality falls with AI involvement, $q | s \sim N(\mu_i - \delta_i s, \sigma_q^2)$, where $\delta_i \geq 0$ measures how much worse the reader expects AI-produced news to be. The byline $\ell \in \{A, B, C\}$ shapes the reader’s belief about production, $\pi_i(\ell) \equiv E_i[s | \ell]$. Reading the article provides a signal $x = q + \varepsilon$ with $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. If the reader does not also infer s from the text, the posterior evaluation is

$$V_i(\ell) = (1 - \kappa) [\mu_i - \delta_i \pi_i(\ell)] + \kappa x, \quad \kappa = \frac{\sigma_q^2}{\sigma_q^2 + \sigma_\varepsilon^2} \in (0, 1), \quad (1)$$

where κ is the weight on the text. The average effect of label ℓ relative to the journalist byline is

$$\tau_\ell \equiv E[V_i(\ell) - V_i(A)] = -(1 - \kappa) E[\delta_i \Delta \pi_i(\ell)] + \gamma_\ell, \quad (2)$$

where $\Delta \pi_i(\ell) = \pi_i(\ell) - \pi_i(A)$ is the shift in beliefs caused by the label and γ_ℓ collects any direct effect of the label, for example a reward for transparency. With $\gamma_\ell = 0$ and δ_i unrelated to $\Delta \pi_i$, the effect is $\tau_\ell = \beta \bar{\Delta \pi}(\ell)$ with $\beta = -(1 - \kappa) \bar{\delta}$.

The model yields four predictions that guide the analysis. (i) A label can affect evaluations only if it changes beliefs about production; the first stage $\bar{\Delta \pi}(\ell)$ is therefore the natural manipulation check. (ii) The penalty per unit of belief falls with the weight on the text, κ . When readers see only a headline, κ is small and a label carries much weight; when they read a full article that they can judge, κ is larger. (iii) If many readers already expect AI to be involved under a journalist byline, $\pi_i(A)$ is high and a label moves beliefs little. (iv) In the cross section, readers who find a text weak may conclude that AI wrote it, so beliefs and evaluations are correlated even if beliefs have no causal effect. The cross-sectional slope of V on π is therefore not an estimate of β . Randomly assigned labels shift π independently of the text signal, so the ratio of the effect on evaluations to the effect on beliefs identifies β under the assumption $\gamma_\ell = 0$. This is the instrumental variables estimator we use in Section 6.7.

3. Experimental design

3.1. Why an experiment

The quantity of interest is causal: how would a reader judge the same article if it carried an AI label instead of a journalist’s name? Let $Y_i(a)$ denote respondent i ’s evaluation if shown byline $a \in \{A, B, C\}$. For any individual we observe only one of these potential outcomes, so the effect must be learned from comparisons across people (Imbens and Rubin, 2015). Observational comparisons would be biased. Outlets that label AI content differ from outlets that do not in topics, style and audience, and readers who encounter labelled articles differ from those who do not. Asking people directly whether an AI label would change their trust measures a stated attitude toward a hypothetical situation. Such answers are informative about opinions (Newman et al., 2024), but they need not predict how people judge an actual article in front of them (Hainmueller et al., 2015; Stantcheva, 2023).

The experiment solves this problem by design. All respondents see the same article, and software assigns the byline at random before the article is shown. Because assignment is random within each country,

$$\{Y_i(A), Y_i(B), Y_i(C), X_i\} \perp T_i \mid m_i, \quad (3)$$

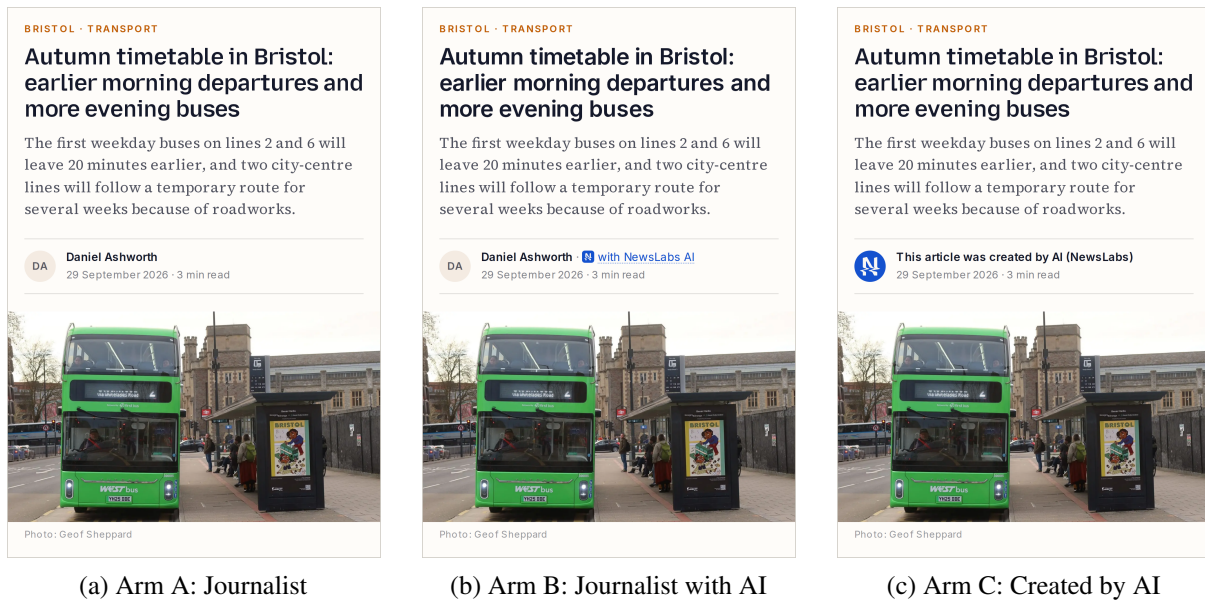
where T_i is the assigned arm, X_i is any characteristic determined before treatment and m_i is the market. The three groups within a market are therefore random draws from the same pool of respondents. They have the same distribution of age, education, attitudes toward AI, reading habits and every other trait, observed or not, up to sampling variation. Any systematic difference in outcomes between the groups can only be caused by the byline, and the difference in mean outcomes identifies the average treatment effect, $E[Y_i \mid T_i = B, m_i] - E[Y_i \mid T_i = A, m_i] = E[Y_i(B) - Y_i(A) \mid m_i]$ (Athey and Imbens, 2017).

The credibility of this argument rests on four features of the design, each of which we can document. First, assignment was made by the survey software, not by respondents or researchers, and it could not respond to anything the respondent did after the article was shown. Second, the byline was the only element that differed: the text, photograph, layout, date and reading time were identical, and the AI provider’s name appeared nowhere else before or during the article. Third, randomization has a testable implication: characteristics measured before treatment should be balanced across arms. Section 4.2 shows that they are broadly balanced, in means and in full distributions. Fourth, outcomes were measured in the same way in every arm, attrition after assignment was very low, and it did not differ by arm. Covariates are therefore not needed to remove bias; we use them only to improve precision and to check the randomization.

3.2. Stimulus and treatments

We built a dedicated mobile web application for the study. After a short introduction, each respondent saw a news article page in a standard layout: a section label, headline, standfirst, byline with the date of participation and an estimated reading time, a full-width photograph of a local city bus, four paragraphs of text, a box with the timetable changes and a cited source. Figure 1 shows the top of the page in the three arms; Appendix B reproduces the full text.

Figure 1. The stimulus in the three arms (United Kingdom version, mobile view)



Notes: Screenshots of the upper part of the article page. The rest of the page (text, timetable box and source) was identical in all arms. In arm B, tapping “with NewsLabs AI” opened a short card describing the assistant (Appendix Figure A2). Other markets used the same layout in the local language, with a local city, photograph and journalist name.

The three bylines were:

- **Arm A (control):** the journalist’s name only, for example “Daniel Ashworth”.
- **Arm B (journalist with AI):** “Daniel Ashworth · with NewsLabs AI”, with a small logo before the words “with NewsLabs AI”. Tapping the label opened a card reading “An artificial intelligence assistant that helps journalists with their work on articles.”
- **Arm C (created by AI):** a logo and the sentence “This article was created by AI (NewsLabs)”, with no journalist.

The article reports minor changes to a city’s autumn bus timetable: earlier first departures on two lines, a shorter evening interval on a third, a temporary route because of roadworks, and unchanged fares. We chose this story for three reasons. It is the kind of routine service news for which newsrooms are most likely to use AI first. It is politically neutral, so partisan reactions do not mix with reactions to the label. And every city has buses, so the same story works in every market. The story is fictional but set in a real city in each market (Bristol, Columbus, Split, Verona, Münster, Poitiers and Valladolid), with a photograph of a local bus, a common local name for the journalist, and local spelling and time formats.

3.3. Markets, sample and fieldwork

Respondents were recruited through an online research panel that is widely used in social science (Peer et al., 2022). Each market was a separate study open to adults resident in that country and fluent in the survey language, and people who took part in one market’s study were excluded from the others. Respondents were paid a fixed amount for completing the survey, independent of their answers. Fieldwork took place in September and October 2026. Each market has about 100 respondents, except Croatia, where the panel is smaller and the sample has 72 respondents. The survey was available in English (UK

and US versions), Croatian, Italian, German, French and Spanish.

3.4. Randomization

Assignment was blocked by market. When a new respondent started, the software assigned the arm with the fewest respondents in that market so far and broke ties at random. This rule belongs to the family of sequential designs that force balance (Efron, 1971). Because arrivals were processed in sequence, it is equivalent to randomizing within consecutive blocks of three respondents in each market, a permuted block design (Rosenberger and Lachin, 2016). Of the 229 blocks, 223 contain exactly one respondent in each arm; the remaining six are the incomplete last blocks in six of the markets. Respondents had no way to influence or foresee their assignment. A respondent who reloaded the page kept the original arm, and the application did not allow returning to the article once the questions about it had started.

3.5. Questionnaire and outcomes

Table 1 lists the questions in the order shown and explains the role of each. The sequence follows three rules. The only covariate collected in the survey, the frequency of reading news, is measured before the article, so it cannot be affected by treatment. The main outcomes are measured immediately after the article and before any question that mentions the author, so that the recall question cannot draw attention to the byline. General attitudes come last, because they are measured after treatment and may be affected by it. Screenshots of all screens are in Appendix Figure A1. The median respondent completed the survey in 142 seconds and spent 62 seconds on the article page.

The three impression items in Q4 measure the core of message credibility and perceived quality used in research on news evaluation (Appelman and Sundar, 2016; Graefe et al., 2018). We combine them into one index to avoid testing the main hypothesis three times (Kling et al., 2007; Anderson, 2008); the items are highly correlated (Cronbach's $\alpha = 0.84$). The reading frequency in Q1 lets us check whether regular readers of news portals, who are used to bylines and to the norms of a newsroom, react differently from people who rarely read news. The reading intention in Q5 is closer to behavior than an attitude scale and measures possible damage directly: if a label hurts, readers should be more likely to skip the article or leave the site. The recall question in Q6 measures whether respondents consciously noticed the byline, while the production belief in Q8 measures whether the label changed what they think, whether or not they remember it. In the model of Section 2.2, Q8 is the first stage. Q7 checks that respondents read the text. Q9 measures a general attitude; because it is asked after treatment, we analyze it as a secondary outcome and never use it as a control.

3.6. Covariates

The panel provides profile data for each respondent, which we merged using an anonymous identifier. The same core set of variables is available in all seven markets: age, sex, highest completed education, self-placed socioeconomic status on a ladder from 1 to 10, frequency of AI use at work, student status, employment status, country of birth, first language and the number of studies previously completed on the panel (a measure of survey experience). We add news reading frequency (Q1) and the type of device, which the application recorded from the browser before treatment. Personal income is available in six

markets but not in Croatia; we convert the reported income bracket to a percentile within each country. Appendix Table [A1](#) reports the availability of each covariate by market. Missing values are rare outside income; we replace them with the sample mean and add an indicator for missingness, which keeps every respondent in the analysis and is valid because the covariates are fixed before treatment ([Lin, 2013](#)).

Table 1. Survey instrument: questions, response options and role in the analysis

	Question (English version)	Response options	Role
Q1	How often do you read news online (on news sites, in apps, or on social media)?	Five options, from “Several times a day” to “Almost never”	Pre-treatment covariate and moderator: separates regular readers of news portals from people who rarely read news and may care less about who wrote an article
Q2	Scene: “Imagine you live in [city]. You take the city bus almost every day. One morning, you come across this article on a local news site.”	None	Makes the story personally relevant, as for a real local reader
Q3	Article page (stimulus). The continue button appears after 5 seconds.	None	Treatment: byline A, B or C
Q4	“Based on your impression of the article you just read”: How good is this article overall? How credible does it seem to you? How professionally written does it seem?	Three scales from 1 to 7, labelled at the ends	Primary outcome: credibility index (mean of the three standardized items)
Q5	If you came across an article like this on a news site you read regularly, what would you most likely do?	Read it to the end; skim it; skip it and look at something else; leave the site	Secondary outcome: direct measure of damage; a harmful label would raise the share who skip the article or leave the site in arms B and C
Q6	How was the author of the article shown? (Please answer from memory.)	Only a journalist’s name; a journalist’s name with the name of a tool or company; no journalist, but a note that the article was created by AI; I don’t remember	Manipulation check: conscious recall of the byline
Q7	According to the article, how much earlier will the first weekday bus on line 2 leave?	10, 20 (correct) or 30 minutes; the article does not say	Attention check; used only in robustness tests
Q8	What do you think: how was the article you just read produced?	Written by a human without AI; mostly by a human with the help of AI; mostly by AI, checked by a human; entirely by AI; I can’t tell	Mechanism: belief about AI involvement
Q9	News can be produced in different ways. Which would you generally trust most? (Introduced as a question about general views.)	Human without AI; human with the help of AI; AI checked by a human (order of these three rotated); it makes no difference	Secondary outcome: general attitude, measured after treatment

Notes: One question or block per screen, in the order listed. Respondents could not go back to earlier screens. Demographic characteristics were not asked in the survey; they come from the panel’s profile data (Section 3.6). At the end, all respondents were told that the article had been written for the study, that it did not report real events, and that versions differed only in how the author was credited.

4. Data and validity of the design

4.1. Sample

Table 2 describes the 679 respondents. They are on average 36 years old, 41 percent are women, 62 percent hold a tertiary degree, 45 percent use AI at work every day and 82 percent read news online at least daily; 80 percent took part on a mobile phone. The samples are not representative of national populations. Like most online panels, they are younger, more educated and more familiar with AI than the average adult. This matters for external validity, which we discuss in Section 7, but not for the comparison between arms.

Ratings are high overall. The mean credibility rating is about 6 on the scale from 1 to 7, and 40 percent of the control arm gave the maximum score. They also differ across countries: the credibility index is highest in the United States and lowest in Croatia, a gap of 0.56 points on the 1 to 7 scale, or 0.61 SD. A notable baseline fact appears in the last column: in every market, at least half of the respondents who saw only a journalist's name believed that AI had been involved in producing the article (62 percent overall).

4.2. Balance of pre-treatment characteristics

If randomization worked, the three arms should look alike before treatment. Figure 2 plots, for each of the 13 pre-treatment covariates, the difference in means between each label arm and the control arm, divided by the pooled standard deviation. This normalized difference does not depend on sample size; values below 0.25 in absolute value are conventionally considered small enough for regression methods to adjust reliably (Imbens and Wooldridge, 2009). All 26 differences are below this threshold (the largest is 0.24), and 16 are below 0.1.

Balance in means does not guarantee that the groups are comparable along the whole distribution, which is the stronger implication of randomization. Figure 3 therefore compares the full cumulative distributions of the six non-binary covariates across arms. The three distributions lie almost on top of each other for age, social status, news reading frequency, AI use at work and panel experience. The k -sample Anderson-Darling test (Scholz and Stephens, 1987), which compares entire distributions, does not reject equality for any of these five variables. The exception is within-country income, which is somewhat higher in arm C ($p = 0.02$).

Appendix Table A2 reports the corresponding tests for all covariates. Two of the 13 covariates differ across arms at the 5 percent level: arm C has fewer women (34 percent against 44 percent in arm A) and a higher average income percentile. With 13 covariates and two comparisons each, one or two differences of this size are expected by chance. Joint tests do not reject balance at the 5 percent level, although the margin is not large: a likelihood ratio test from a multinomial logit of the arm on all covariates and market fixed effects gives $\chi^2(40) = 53.0$ ($p = 0.08$), and a randomization test of an omnibus statistic that sums the covariate-by-covariate F statistics gives $p = 0.06$. Appendix Table A3 repeats the omnibus test within each market, where each arm has only about 33 respondents. Balance is not rejected in six of the seven markets; in the United States, respondents in arm B are older and those in arm C are less often students ($p = 0.007$), a result that is not unusual among seven tests. Because chance imbalances of this

Table 2. Sample characteristics and outcome means by market

Market	N	Pre-treatment characteristics							Outcomes (all arms)				
		Age (yrs)	Female (%)	Tertiary (%)	SES (1-10)	AI daily (%)	News daily (%)	Mobile (%)	Qual. (1-7)	Cred. (1-7)	Prof. (1-7)	Read end (%)	AI inv. in A (%)
United Kingdom	101	37.7	53	65	5.2	32	74	72	5.46	5.85	5.52	32	67
United States	103	39.2	52	62	5.1	37	75	82	5.78	6.12	6.04	48	69
Croatia	72	35.6	46	62	5.8	39	77	83	5.28	5.58	5.39	42	78
Italy	101	34.6	36	51	5.5	47	94	87	5.84	6.12	5.67	42	62
Germany	101	34.7	28	61	5.6	54	87	84	5.64	6.07	5.73	45	61
France	101	35.0	36	71	5.5	49	75	83	5.28	5.83	5.33	24	56
Spain	100	35.8	34	64	5.7	54	88	70	5.87	6.12	5.88	37	50
All markets	679	36.1	41	62	5.5	45	82	80	5.61	5.97	5.67	38	62

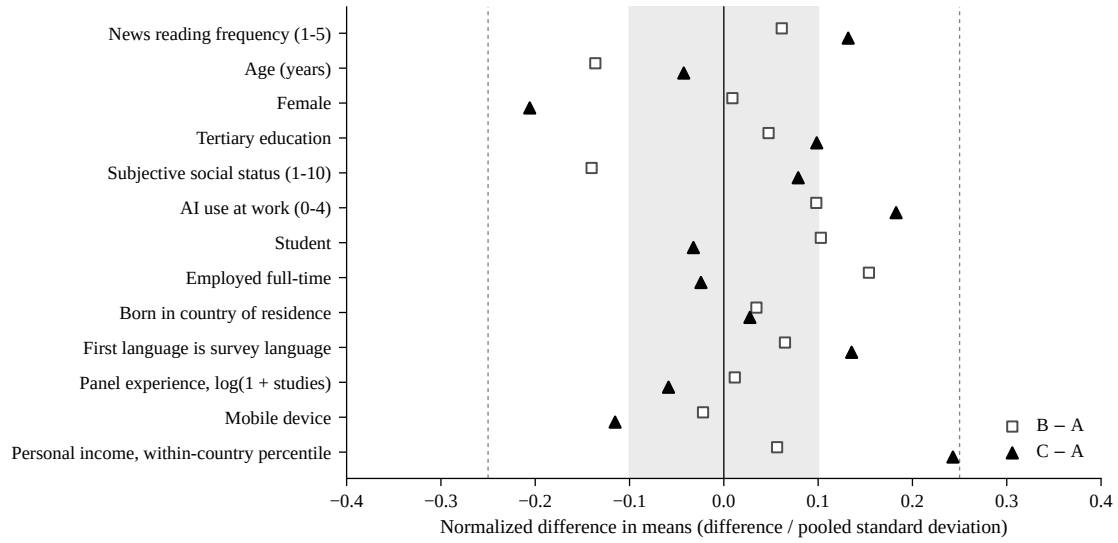
Notes: *N* is the number of respondents assigned to an arm. Means are computed over respondents with a valid value. Age, sex, education, SES, AI use at work and country of birth come from the panel's profile data. *Tertiary*: completed an undergraduate, graduate or doctoral degree. *SES*: self-placed socioeconomic status on a ladder from 1 (lowest) to 10 (highest). *AI daily*: uses AI at work every day or several times a day. *News daily*: reads news online at least once a day (Q1, asked before the article). *Mobile*: completed the survey on a phone, detected from the browser. *Qual.*, *Cred.*, *Prof.*: mean ratings of overall quality, credibility and professional writing on scales from 1 to 7 (Q4). *Read end*: would read an article like this to the end (Q5). *AI inv. in A*: share of respondents in the control arm (journalist byline only) who believed that AI was involved in producing the article (Q8).

kind add noise rather than bias, we report estimates with and without covariate adjustment and show in Section 6.8 that none of the conclusions depends on them, including when we drop the United States or let the covariate coefficients differ by market.

4.3. Attrition, attention and reading time

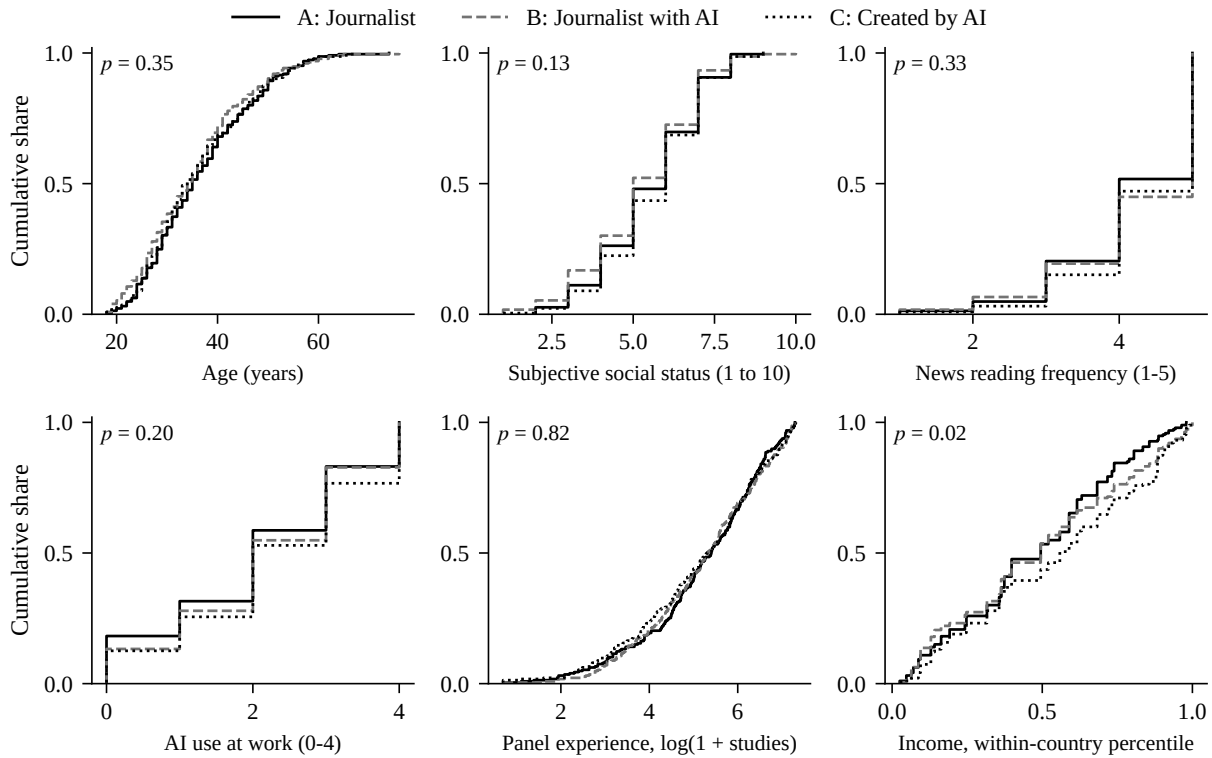
Only 1.3 percent of assigned respondents did not complete the survey, and attrition does not differ by arm ($p = 0.22$). Bounds that trim the arm with higher retention (Lee, 2009) are narrow: $[-0.033, 0.012]$ SD for arm B and -0.074 SD for arm C, where retention is essentially identical to the control arm. In total, 85 percent of respondents answered the attention check correctly, with no difference between arms (84, 85 and 84 percent; $p = 0.97$). Respondents therefore read the article with similar care in all arms, at least as measured by comprehension of its main fact.

Figure 2. Balance of pre-treatment characteristics: normalized differences



Notes: Difference in means between arm B (hollow squares) or arm C (triangles) and arm A, divided by the standard deviation in the pooled sample. The shaded band marks differences smaller than 0.1 in absolute value; the dashed lines mark 0.25, the conventional threshold for a large imbalance (Imbens and Wooldridge, 2009). Appendix Table A2 reports the underlying means and tests.

Figure 3. Balance of pre-treatment characteristics: full distributions by arm



Notes: Empirical cumulative distribution functions by arm. The p -values are from k -sample Anderson-Darling tests of equal distributions across the three arms, computed by permutation. Income is not available for Croatia.

5. Empirical strategy

We estimate intention-to-treat effects with the regression

$$Y_i = \alpha + \tau_B B_i + \tau_C C_i + \mu_{m(i)} + X_i' \beta + \varepsilon_i, \quad (4)$$

where B_i and C_i indicate assignment to arms B and C, μ_m are market fixed effects, which absorb the blocking by market, and X_i contains the 13 pre-treatment covariates and missing-value indicators. We report three versions of the estimate: the simple difference in means, the estimate with market fixed effects only, and the estimate with fixed effects and covariates. The last is our preferred specification because it is the most precise. Standard errors are heteroskedasticity-robust (HC2), which have good properties in samples of this size (MacKinnon and White, 1985). As further checks, we report the fully interacted estimator of Lin (2013), which cannot reduce asymptotic precision, and a post-double-selection lasso that chooses controls from 85 candidates, including squares and market interactions (Belloni et al., 2014).

Primary outcome. Because the study was not pre-registered, we chose the most standard summary measure as the primary outcome and use it throughout: the credibility index, defined as the mean of the three impression items after standardizing each by the control arm, re-standardized to have mean zero and unit standard deviation in the control arm (Kling et al., 2007). Effects are therefore measured in standard deviations of the control group. On the original scale, one control-group standard deviation of the average rating equals 0.92 points, and the control mean is 5.78.

Randomization inference. Conventional standard errors rely on large-sample approximations. As a check that relies only on the design, we compute randomization p -values (Young, 2019; Athey and Imbens, 2017). We reassign arms by permuting the arm labels within the actual randomization blocks (consecutive triplets within each market), re-estimate the effects and compare the observed estimate with this placebo distribution. We use 5,000 permutations for the primary outcome and the main secondary outcomes and 2,000 for the others.

Equivalence tests. A statistically insignificant estimate is not evidence that the effect is small (Abadie, 2020). To make a positive statement, we test the null hypothesis that the effect is at least as large as a stated bound against the alternative that it is smaller, using two one-sided tests (Lakens, 2017). We use bounds of ± 0.2 SD, the conventional threshold for a small effect; on the original scale this equals about 0.18 points. Rejecting both one-sided nulls at the 5 percent level is equivalent to a 90 percent confidence interval that lies entirely within the bounds. We also report the smallest bound for which equivalence can be established.

Multiple testing and power. For the family of five evaluative outcomes (the three impression items, reading to the end and the four-level reading intention) and two arms, we report Holm and Benjamini-Hochberg adjusted p -values (Holm, 1979; Benjamini and Hochberg, 1995). With the realized standard

errors, the minimum detectable effect on the primary outcome at 5 percent significance and 80 percent power is 0.25 SD for arm B and 0.26 SD for arm C.

6. Results

6.1. The AI labels did not reduce credibility

Table 3 reports the effects on the primary outcome. Across all specifications, the estimates are small, negative and far from statistical significance. In the simple comparison of means (column 1), the effect of the “with AI” byline is -0.021 SD and the effect of the “created by AI” byline is -0.074 SD, with randomization p -values of 0.81 and 0.41. With market fixed effects and covariates (column 3), the estimates are -0.050 SD (SE 0.091) and -0.054 SD (SE 0.093). The Lin and post-double-selection estimators give very similar results (columns 4 and 5). The two labels are not different from each other ($p = 0.97$), so Panel B pools them into a single “any AI label” indicator. The pooled effect is -0.052 SD (SE 0.079).

Figure 4 plots the estimates for the index, each of its components and the measures of reading behavior, all in control-group standard deviations. No component shows a significant effect. The largest point estimate is a reduction of about 0.1 points in perceived professional writing in both label arms ($p = 0.26$ and 0.38 ; Table 4). After correction for multiple testing across the five evaluative outcomes and two arms, the smallest adjusted p -value is 1.00 (Holm) and the smallest false discovery rate q -value is 0.98. Ordered logit models, which use the full ordinal distribution of each item, give the same picture (Appendix Table A4).

The null result does not come from opposite changes in different parts of the distribution that cancel out in the mean. Figure 5 shows that the full distributions of the index and of the credibility rating are almost identical in the three arms. Kolmogorov-Smirnov tests do not reject equal distributions of the index between each label arm and the control arm ($p = 0.72$ for arm B and $p = 0.98$ for arm C), and a k -sample Anderson-Darling test on the average rating across the three arms gives $p = 0.89$. The high ratings raise the concern that a ceiling might hide a penalty at the top, but the share of respondents who gave the maximum credibility rating of 7 is also unaffected (Table 8).

6.2. How small is the effect?

The 95 percent confidence intervals for the effect on the index are $[-0.23, 0.13]$ SD for arm B and $[-0.24, 0.13]$ SD for arm C. On the original 1 to 7 scale, they correspond to $[-0.21, 0.12]$ and $[-0.22, 0.12]$ points. For each label separately, the equivalence test against bounds of ± 0.2 SD is at the margin of the 5 percent level ($p = 0.050$ for arm B and $p = 0.057$ for arm C). The two labels do not differ from each other, and pooling them gives a 95 percent interval of $[-0.21, 0.10]$ SD. For the pooled label, the equivalence test rejects effects of 0.2 SD or more ($p = 0.030$), and equivalence holds for any bound of 0.18 SD or more.

Three comparisons help to judge these magnitudes. First, the largest penalty consistent with the pooled estimate, about 0.2 SD or 0.2 points on the 7-point scale, is roughly a third of the difference in average ratings between the most and the least favorable country in our sample (0.61 SD). Second, an

Table 3. Main result: effect of the byline label on the credibility index

	Difference in means (1)	Market fixed effects (2)	Fixed effects and covariates (3)	Lin (2013) estimator (4)	Post-double- selection lasso (5)
<i>Panel A. Separate effects of the two labels</i>					
B: Journalist with AI	−0.021 (0.093) [0.812]	−0.022 (0.091) [0.826]	−0.050 (0.091) [0.598]	−0.028 (0.093)	−0.074 (0.087)
C: Created by AI	−0.074 (0.097) [0.414]	−0.073 (0.095) [0.436]	−0.054 (0.093) [0.566]	−0.048 (0.093)	−0.074 (0.092)
<i>p</i> -value, C = B	0.583	0.588	0.972	0.835	0.999
Equivalence test, $ \tau \geq 0.2$ SD: <i>p</i> (B)			0.050		
Equivalence test, $ \tau \geq 0.2$ SD: <i>p</i> (C)			0.057		
Observations	675	675	675	675	675
Market fixed effects	No	Yes	Yes	Yes	Yes
Pre-treatment covariates	No	No	Yes	Interacted	Lasso-selected
<i>Panel B. Pooled effect of any AI label</i>					
Any AI label (B or C)	−0.047 (0.082)	−0.047 (0.080)	−0.052 (0.079)		
Equivalence test, $ \tau \geq 0.2$ SD: <i>p</i>			0.030		
Observations	675	675	675		

Notes: The dependent variable is the credibility index: the mean of the three standardized impression items (overall quality, credibility, professional writing), standardized to mean 0 and standard deviation 1 in arm A. Effects are relative to arm A (journalist only). HC2 standard errors in parentheses. Randomization inference *p*-values in brackets, from 5,000 permutations of arm labels within randomization blocks. Column (4) interacts the arm indicators with demeaned covariates (Lin, 2013). Column (5) selects controls by lasso from 85 candidate variables (covariates, squares and interactions with market) and always keeps market fixed effects (Belloni et al., 2014). The equivalence test reports the larger of the two one-sided *p*-values against the bounds ± 0.2 SD.

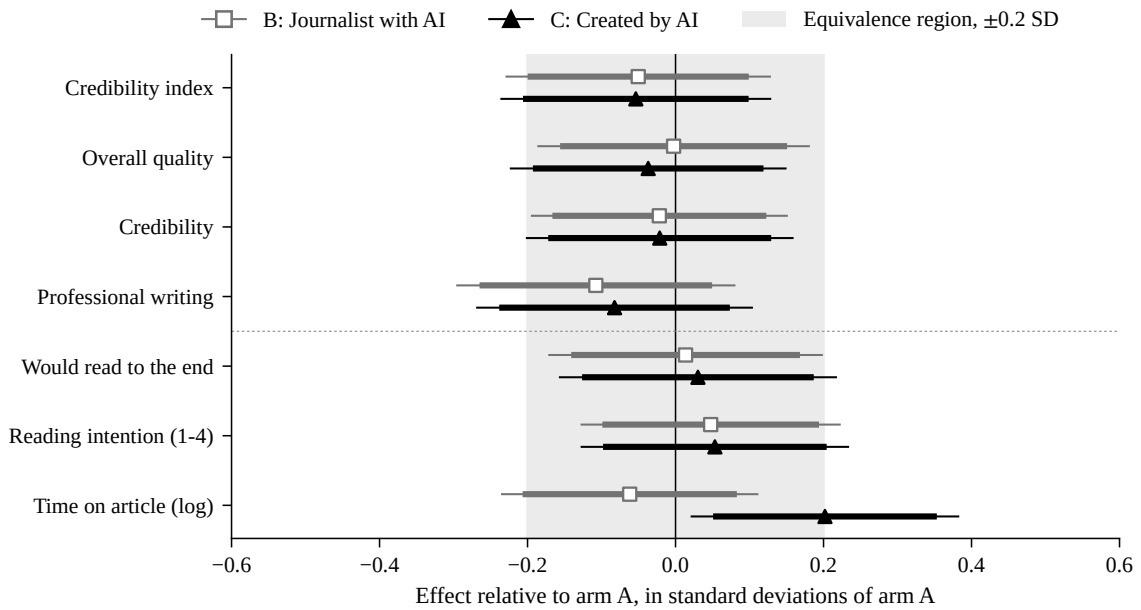
average reduction of 0.2 points is what we would observe if one respondent in five lowered each of the three ratings by one point. Third, earlier experiments report label effects of a similar order: in Toff and Simon (2025), an AI label reduced the rated trustworthiness of a news organization from 5.9 to 5.5 on a scale from 0 to 10, roughly 0.17 SD by our calculation, and in Wittenberg et al. (2025), an “AI-generated” label reduced belief in misleading images by 0.14 to 0.18 SD. Our design cannot exclude penalties of that size, but our point estimates are about a third as large.

6.3. No sign of damage to reading

Question Q5 measures the outcome that matters most for a news portal: if a label hurt, more readers in arms B and C would skip the article or leave the site. This did not happen. The share who would leave the site is 2.2 percent in the control arm and 0.4 percent in each label arm, and the share who would skip the article is 4.4, 4.8 and 5.8 percent. The share who would read an article like this to the end is 38 percent in the control arm and in arm B and 39 percent in arm C; the estimated effects are 0.7 and 1.5 percentage points, with standard errors of about 4.6 points (Table 4, column 4). The full distribution of intentions is shown in Appendix Figure A3.

One behavioral difference does appear. Respondents in arm C spent somewhat longer on the article page: the median was 66 seconds, against 60 seconds in the control arm. In the regression with covariates, time on the page in arm C was about 14 percent longer than in arm A ($p = 0.03$; randomization $p = 0.03$)

Figure 4. Effects of the labels on evaluations and reading behavior



Notes: Estimates of equation (4) with market fixed effects and covariates. All outcomes are standardized by the mean and standard deviation of arm A. Thick bars are 90 percent confidence intervals and thin lines 95 percent confidence intervals (HC2). An estimate whose 90 percent interval lies inside the shaded region is statistically equivalent to zero at the 5 percent level for bounds of ± 0.2 SD. Outcomes above the dotted line are the evaluation measures (Q4); outcomes below it are reading measures (Q5 and time on the article page).

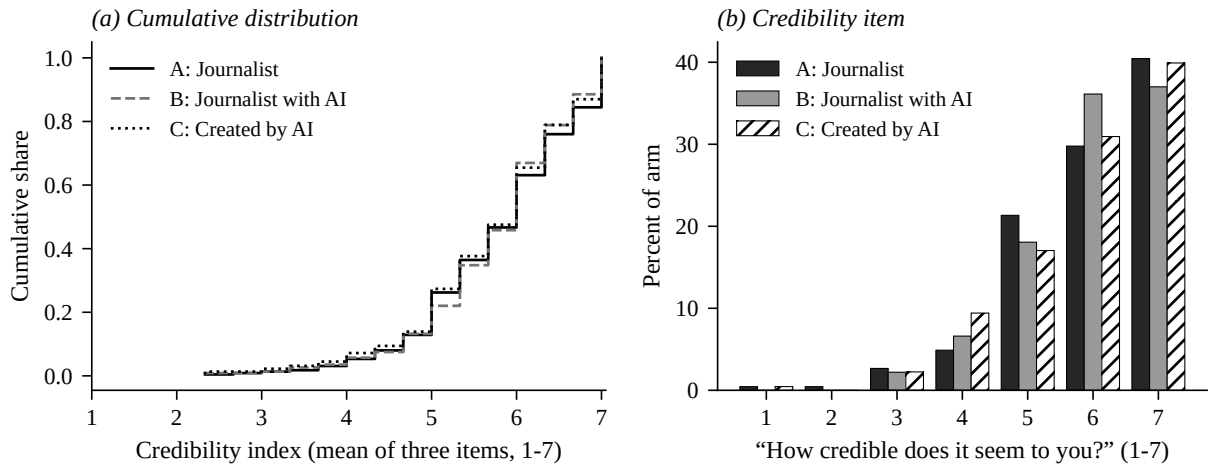
and about 19 percent longer than in arm B ($p = 0.003$). Without covariates, the difference to arm A is 12 percent ($p = 0.06$). Since reading time is one of several secondary outcomes, we treat this result as exploratory. It suggests that a clear “created by AI” label leads some readers to read more carefully rather than to disengage, and that this extra scrutiny did not lower their evaluations.

6.4. Readers assume AI involvement by default

The second main result concerns what readers believe when there is no label at all. In the control arm, where the byline showed only a journalist’s name, 62 percent of respondents believed that AI had been involved in producing the article, and only 18 percent thought it was written by a human without AI (Figure 6, panel b, first row). This prior is widespread. It ranges from 50 percent in Spain to 78 percent in Croatia (Table 2, last column), and it is similar across groups of readers (Figure 8, panel c): 65 percent among respondents who never use AI at work and 62 percent among daily users, and between 61 and 65 percent in each age group. It is somewhat higher among respondents without a university degree (73 percent) and among those who read news less than daily (73 percent). For most readers, then, AI assistance is a normal assumption about how news is made today, and a label that discloses it mostly confirms what they already expect. This is prediction (iii) of the model in Section 2.2.

6.5. Most readers did not know which label they saw

The third main result is that the byline label largely went unnoticed. Figure 6, panel (a), shows the answers to the recall question, asked from memory after the impression questions. Overall, 66 percent of respondents said they did not remember how the author was shown: 65 percent in the control arm,

Figure 5. Distribution of evaluations by arm

Notes: Panel (a): empirical cumulative distribution of the credibility index on the original scale (mean of the three items, 1 to 7). Panel (b): distribution of answers to “How credible does it seem to you?” (1 = not at all credible, 7 = completely credible).

Table 4. Effects on the components of the index and on reading behavior

	Impression items (1 to 7)			Reading behavior		
	Overall quality (1)	Credibility (2)	Professional writing (3)	Would read to the end (4)	Reading intention (1 to 4) (5)	log time on article page (6)
B: Journalist with AI	−0.003 (0.092) [0.975]	−0.024 (0.098) [0.802]	−0.116 (0.103) [0.246]	0.007 (0.046) [0.895]	0.031 (0.058) [0.578]	−0.041 (0.058) [0.498]
C: Created by AI	−0.036 (0.093) [0.699]	−0.024 (0.102) [0.822]	−0.089 (0.102) [0.353]	0.015 (0.046) [0.748]	0.035 (0.060) [0.553]	0.133 (0.060) [0.028]
<i>p</i> -value, C = B	0.726	0.995	0.799	0.862	0.949	0.003
Mean in arm A	5.63	5.97	5.73	0.38	3.29	4.05
Observations	675	675	675	675	675	676

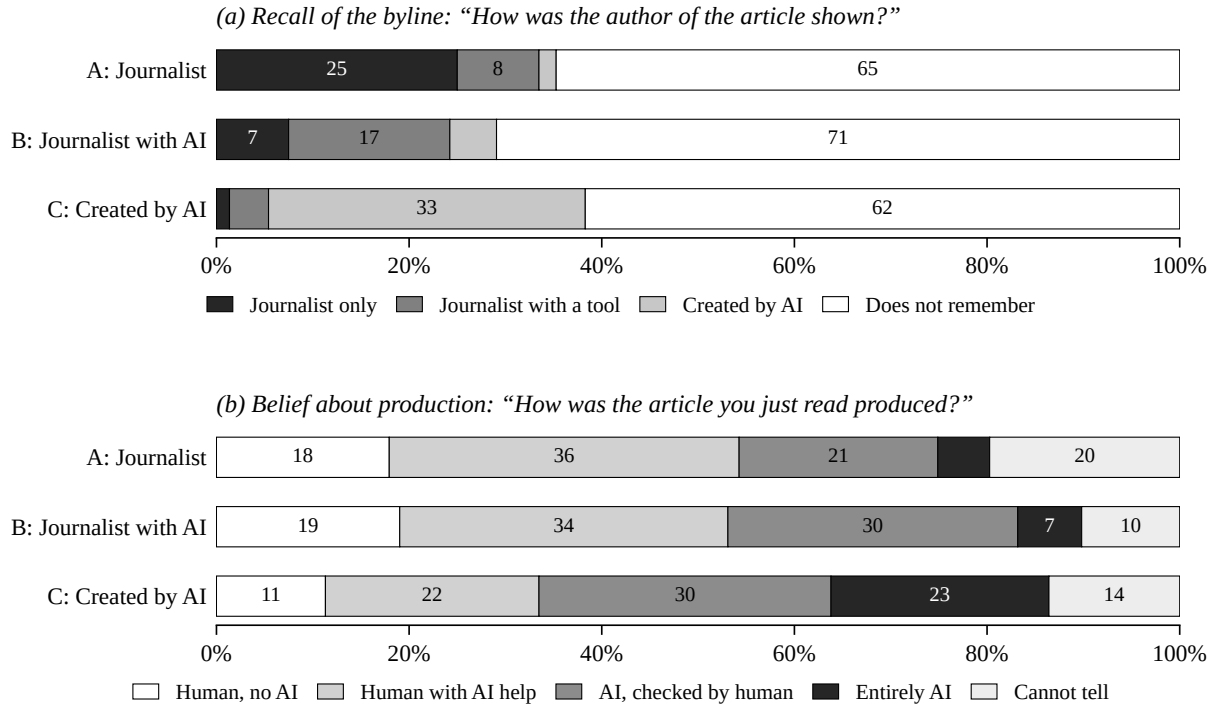
Notes: OLS estimates with market fixed effects and pre-treatment covariates; HC2 standard errors in parentheses; randomization inference *p*-values in brackets, from the same specification. Columns (1) to (3) are in points on the original 1 to 7 scale. Column (4) is a linear probability model. Column (5) codes leave the site = 1, skip = 2, skim = 3 and read to the end = 4. Column (6) is the log of seconds spent on the article page; coefficients are approximately proportional changes.

71 percent in arm B and 62 percent in arm C. Among those who did remember, the answers match the assigned arm. In arm C, 33 percent recalled a note that the article was created by AI, against 2 percent in the control arm. In arm B, 17 percent recalled a journalist’s name with a tool or company, against 8 percent in the control arm (Table 6, columns 1 to 3). The soft label in arm B was the hardest to notice, and no respondent in arm B tapped it to read the explanation of the assistant. Bylines are a small part of a news page, and most readers do not read them closely. We placed the label where a publisher would, so this low attention is part of what we measure.

6.6. Heterogeneity across countries and readers

Figure 7 shows the effect on the index separately for each country. With about 33 respondents per arm, each estimate is imprecise, but the estimates scatter around zero and none is significant at the 5 percent level. A joint test that the effects are equal in all seven markets does not reject ($p = 0.48$). Treating the

Figure 6. Recall of the byline and beliefs about production, by arm



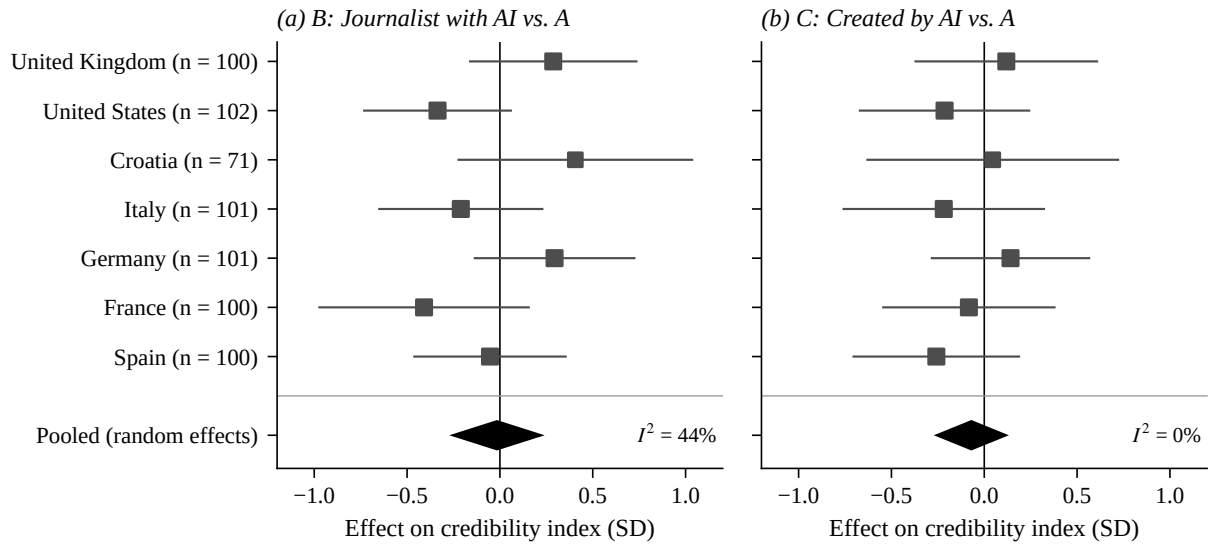
Notes: Percent of respondents in each arm choosing each answer. Panel (a): Q6, answered from memory after the impression questions. Panel (b): Q8. Numbers below 5.5 percent are not printed.

seven countries as seven replications of the same experiment and pooling them with a random-effects model (DerSimonian and Laird, 1986; Meager, 2019) gives pooled effects close to zero; the estimated between-country variance is zero for arm C and moderate but not significant for arm B (Appendix Table A6). The effect in the EU markets, where the disclosure rule applies, is not significantly different from the effect in the UK and the US (Table 5). In Croatia, the market with the lowest stated comfort with AI-produced news in the Reuters survey (Newman et al., 2025), both point estimates are positive.

Figure 8 and Table 5 examine differences across readers, using characteristics measured before treatment. The comparison yields a clear pattern: general attitudes toward AI-made news differ strongly across readers, but their reaction to a labelled article does not.

General attitudes depend on experience with AI. Among respondents who never use AI at work, 72 percent say they generally trust news written by a human without AI most; among daily users, only 41 percent do (panel b). Each step on the five-level scale of AI use at work lowers this share by about 8 percentage points ($p < 0.001$). Older respondents are slightly less likely than younger ones to prefer human-only news (48 percent at age 40 and older against 55 percent below 30; $p = 0.01$ for a linear age term).

Reactions to the label do not follow these attitudes (panel a). The effect of the label on the credibility index does not change with AI use at work (interaction with the five-level scale 0.03 SD per step, $p = 0.62$), with age ($p = 0.84$ for a linear term) or with how often respondents read news ($p = 0.40$). The only group with a sizable point estimate is the small group that never uses AI at work (-0.33 SD, SE 0.25, $n = 99$), and this estimate is not significant. Women react somewhat less favorably than men (-0.21 against $+0.07$ SD) and respondents with a university degree somewhat less favorably than those without (-0.13

Figure 7. Effects on the credibility index by country

Notes: Difference in means of the credibility index (in standard deviations of the pooled control arm) between each label arm and arm A within each country, with 95 percent confidence intervals (HC2). Marker size is proportional to the sample size. The diamonds show the pooled random-effects estimate and its 95 percent confidence interval (DerSimonian and Laird, 1986); I^2 is the share of the variation in estimates due to heterogeneity across countries.

Table 5. Heterogeneity of effects on the credibility index by reader characteristics

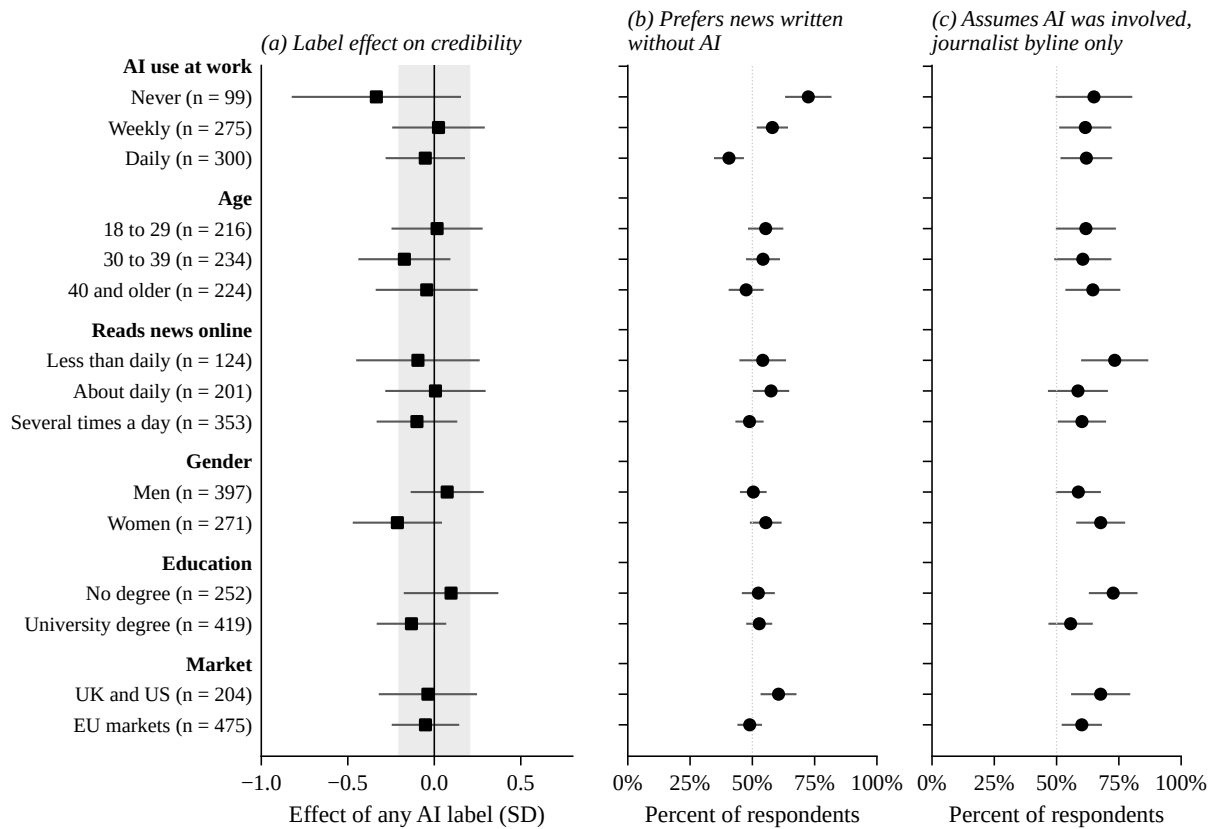
Moderator	Share (%)	B $\times$ moderator	(SE)	C $\times$ moderator	(SE)	Joint p	BH q
EU market (vs. UK and US)	70	0.007	(0.190)	-0.036	(0.205)	0.974	0.974
Uses AI at work daily	45	-0.028	(0.183)	0.063	(0.194)	0.891	0.974
Age at or above median	50	0.039	(0.183)	-0.054	(0.191)	0.885	0.974
Reads news several times a day	52	0.005	(0.180)	-0.166	(0.190)	0.595	0.974
Female	41	-0.303	(0.185)	-0.292	(0.208)	0.198	0.840
Tertiary education	62	-0.166	(0.190)	-0.280	(0.197)	0.360	0.840
Mobile device	80	-0.130	(0.224)	0.235	(0.243)	0.297	0.840

Notes: Each row is a separate regression of the credibility index on the arm indicators, the moderator, their interactions and market fixed effects. The columns report the share of the sample with the moderator equal to one, the interaction coefficients with HC2 standard errors, the p -value of a joint test that both interactions are zero, and the Benjamini-Hochberg adjusted q -value across the seven moderators. All moderators are measured before treatment.

against +0.10 SD), but none of the interactions in Table 5 is significant after correction for multiple testing (smallest joint $p = 0.20$, $q \geq 0.84$). Attention to the byline also differs: respondents aged 40 and older (75 percent) and respondents in the UK and the US (78 percent) were the most likely to say they did not remember how the author was shown. We conclude that the average null does not hide large effects of opposite sign in groups we can observe, even among readers whose stated attitudes toward AI-made news are the least favorable.

The labels may even have softened general attitudes. The last question asked which way of producing news respondents trust most in general. After reading an article labelled “created by AI”, 46 percent preferred news written by humans without AI, against 55 percent in the control arm (9 percentage points, $p = 0.05$ without covariates and $p = 0.08$ with covariates), and the share preferring AI-assisted forms of production was higher by a similar amount (Appendix Table A5). These estimates are marginal and not corrected for multiple testing. If they hold up, they suggest that exposure to a competent article openly labelled as AI-produced slightly softens general skepticism, which is the opposite of the effect publishers

Figure 8. Label effects, general attitudes and prior beliefs by reader characteristics



Notes: Panel (a): effect of any AI label (arms B and C pooled) on the credibility index within each group, from a regression with market fixed effects, with 95 percent confidence intervals (HC2); the shaded band marks ± 0.2 SD. Panel (b): share of respondents in all arms who would generally trust news written by a human without AI most (Q9). Panel (c): share of respondents in the control arm (journalist byline only) who believed that AI was involved in producing the article (Q8). All characteristics are measured before treatment.

fear and in line with evidence that mandatory labels need not harden attitudes (Kolodinsky and Lusk, 2018).

6.7. The labels shifted stated beliefs, but not ratings

The labels did change what some readers believed about how the article was made. The share who believed that the article was produced mostly or entirely by AI was 26 percent in the control arm, against 37 percent in arm B (11 percentage points, $p = 0.01$; randomization $p = 0.02$) and 53 percent in arm C (27 percentage points, $p < 0.001$). The share who believed it was produced entirely by AI was 5 percent in the control arm and 23 percent in arm C (Table 6, columns 4 to 7). On the scale from 0 to 3, the perceived AI share was 0.56 points higher in arm C than in the control arm.

These belief shifts did not carry over to ratings. Looking only at correlations would suggest the opposite. Panel (a) of Figure 9 shows the average credibility index by the respondent's belief about how the article was produced, pooling all arms. Ratings fall steadily with perceived AI involvement: respondents who believed that the article was written entirely by AI rated it 0.83 points lower on the 1 to 7 scale than those who believed it was written by a human without AI (0.81 SD after adjusting for market and covariates). The OLS slope is -0.271 SD per step on the scale of perceived AI share (Table 7).

This gradient overstates the causal effect of a label. In arm C, the perceived AI share was 0.56 points

Table 6. Effects of the labels on recall of the byline and on beliefs about production

	Recall of the byline (Q6)			Belief about production (Q8)			
	Correct recall (1)	Does not remember (2)	Recalls “created by AI” (3)	Any AI involvement (4)	Mostly or entirely AI (5)	Entirely AI (6)	AI share (0 to 3) (7)
B: Journalist with AI	−0.087 (0.039) [0.056]	0.069 (0.044) [0.204]	0.038 (0.019) [0.312]	0.074 (0.045) [0.045]	0.111 (0.044) [0.016]	0.018 (0.022) [0.615]	0.121 (0.089) [0.264]
C: Created by AI	0.075 (0.043) [0.053]	−0.028 (0.046) [0.465]	0.314 (0.032) [<0.001]	0.134 (0.044) [0.004]	0.265 (0.045) [<0.001]	0.176 (0.032) [<0.001]	0.563 (0.098) [<0.001]
<i>p</i> -value, C = B	<0.001	0.027	<0.001	0.159	0.001	<0.001	<0.001
Mean in arm A	0.25	0.65	0.02	0.62	0.26	0.05	1.17
Observations	673	673	673	670	670	670	573

Notes: OLS estimates of equation (4) with market fixed effects and pre-treatment covariates. Columns (1) to (6) are linear probability models. Column (7) codes the belief from 0 (written by a human without AI) to 3 (entirely by AI) and excludes respondents who answered “I can’t tell”. HC2 standard errors in parentheses. Randomization inference *p*-values in brackets, from 2,000 or more permutations of arm labels within randomization blocks (specification with market fixed effects).

Table 7. Beliefs about AI production and evaluations: OLS and 2SLS

Endogenous belief (Q8)	OLS		2SLS		Anderson-Rubin	First-stage	<i>N</i>
	Coef.	(SE)	Coef.	(SE)	95% set	<i>F</i>	
Perceived AI share (0 to 3)	−0.271	(0.044)	−0.219	(0.163)	[−0.66, 0.23]	37.4	573
Mostly or entirely AI (0/1)	−0.407	(0.076)	−0.169	(0.339)	[−1.04, 0.80]	35.7	670
Any AI involvement (0/1)	−0.297	(0.081)	−0.379	(0.674)	[−3.29, 2.37]	9.5	670

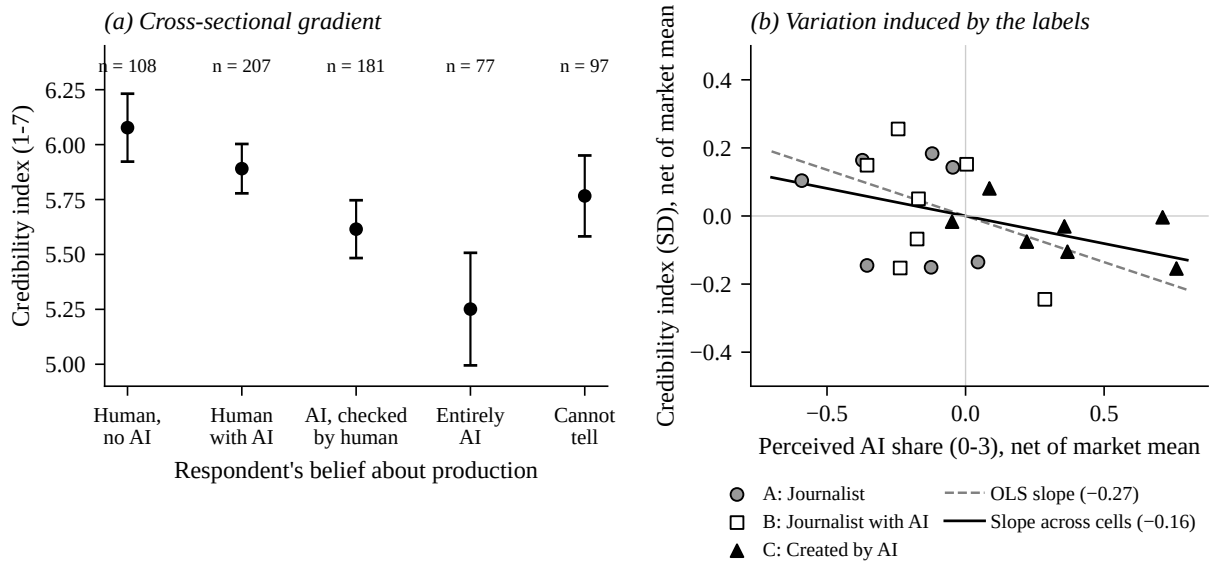
Notes: The dependent variable is the credibility index (SD units). In the 2SLS models, the belief is instrumented with the indicators for arms B and C. All models include market fixed effects and pre-treatment covariates. HC2 standard errors for OLS and heteroskedasticity-robust standard errors for 2SLS. The first row excludes respondents who answered “I can’t tell”. The instruments are weak for “any AI involvement”, so the last row is not informative.

higher than in the control arm; if the cross-sectional slope were causal, the label should have lowered ratings by about $0.56 \times -0.271 \approx -0.15$ SD. The estimated effect is -0.054 SD, a third of that. Panel (b) of Figure 9 shows the same point graphically. It plots, for each market and arm, the mean belief against the mean rating, net of market means, so that the only source of variation is the random label. The slope across these 21 cells (-0.16) is flatter than the individual-level slope.

Table 7 makes this comparison formal. We estimate the effect of the perceived AI share on the credibility index by two-stage least squares, using the arm indicators as instruments for beliefs. In the language of the model in Section 2.2, this estimates $\beta = -(1 - \kappa)\bar{\delta}$ under the assumption that labels affect evaluations only through beliefs ($\gamma_\ell = 0$). The first stage is strong ($F = 37$). The 2SLS estimate is -0.219 SD per step (SE 0.163), about a fifth smaller than the OLS estimate of -0.271 . Because the labels move beliefs only modestly, the 2SLS estimate is imprecise: the Anderson-Rubin confidence set, which is valid even if instruments are weak (Andrews et al., 2019), runs from -0.66 to 0.23 , so we cannot reject either zero or the OLS value. The same pattern holds when the belief is coded as a binary indicator for “mostly or entirely AI”, where the 2SLS estimate is less than half of the OLS estimate.

We read the evidence as follows. Much of the raw association between believing that AI wrote an article and rating it lower likely reflects inference in the other direction: readers who find a text weak

Figure 9. Beliefs about production and evaluations: cross-sectional and experimental variation



Notes: Panel (a): mean credibility index (1 to 7) by belief about production (Q8), all arms pooled, with 95 percent confidence intervals. This relationship is not causal. Panel (b): each point is one of the 21 market-by-arm cells. Both axes show the cell mean minus the market mean, so that only variation created by the random labels remains. The dashed line has the slope of the individual-level OLS regression of the index on the perceived AI share (first row of Table 7); the solid line is the weighted least squares fit across cells, the visual analogue of the instrumental variables estimate.

conclude that a machine produced it, as prediction (iv) of the model suggests. Whatever the true causal effect of beliefs, the label on its own moves beliefs too little to produce a large change in ratings, because many readers do not process the byline and many already assume AI involvement. Two caveats apply. The 2SLS estimate cannot rule out that the cross-sectional slope is causal, and the exclusion restriction would fail if a label also had a direct effect, for example if readers rewarded transparency. In that case a negative belief effect would be partly offset by a positive direct effect, and the small reduced-form effect would still be the relevant quantity for a publisher deciding whether to disclose.

6.8. Robustness

Table 8 shows that the main result does not depend on the sample or the estimator. The estimates for both labels stay between -0.11 and $+0.03$ SD and are never close to significance when we: use the simple difference in means; keep only respondents who passed the attention check (a post-treatment variable, but one that is balanced across arms); drop respondents who spent less than 20 seconds on the article, who reloaded the page or whose total time was in the top or bottom percentile; keep only respondents born in the country who answered in their first language; keep only respondents with a complete profile; drop Croatia, the United States, or both non-EU markets; allow the effect of each covariate to differ by market; keep only the Croatian respondents from the first fieldwork wave; use the raw average of the items or their first principal component instead of the index; or use the Lin and lasso estimators. The share giving the maximum credibility rating is also unaffected.

Table 8. Robustness of the main result

Specification	N	B: Journalist with AI		C: Created by AI	
		Coef.	(SE)	Coef.	(SE)
Baseline: market fixed effects and covariates	675	−0.050	(0.091)	−0.054	(0.093)
Difference in means	675	−0.021	(0.093)	−0.074	(0.097)
Passed the attention check	567	−0.022	(0.099)	0.026	(0.099)
At least 20 seconds on the article	634	−0.032	(0.093)	−0.046	(0.096)
No page reloads	654	−0.054	(0.092)	−0.054	(0.094)
Trim 1% shortest and longest durations	656	−0.054	(0.091)	−0.054	(0.094)
Born in country, survey in first language	519	−0.050	(0.105)	−0.067	(0.111)
Complete demographic profile	647	−0.040	(0.092)	−0.057	(0.094)
Excluding Croatia	604	−0.093	(0.093)	−0.054	(0.096)
Excluding the United States and United Kingdom	473	−0.073	(0.112)	−0.064	(0.111)
Excluding the United States	573	−0.019	(0.100)	−0.040	(0.101)
Covariates interacted with market	675	−0.104	(0.091)	−0.105	(0.100)
Croatia: first fieldwork wave only	651	−0.040	(0.091)	−0.037	(0.093)
Unweighted mean of raw items (1 to 7)	675	−0.048	(0.084)	−0.050	(0.086)
First principal component of the items	675	−0.051	(0.091)	−0.054	(0.093)
Top box: credibility rated 7 of 7 (LPM)	675	−0.038	(0.045)	0.008	(0.045)
Lin (2013) interacted estimator	675	−0.028	(0.093)	−0.048	(0.093)
Post-double-selection lasso	675	−0.074	(0.087)	−0.074	(0.092)

Notes: The dependent variable is the credibility index in control-group standard deviations, except in three rows: the unweighted mean of the raw items is in points on the 1 to 7 scale; the first principal component is standardized on arm A; the top-box row is a linear probability model for rating credibility 7 out of 7. All rows include market fixed effects and pre-treatment covariates unless stated otherwise. HC2 standard errors in parentheses. “First wave” keeps the Croatian respondents collected in the first fieldwork wave, together with the other markets.

7. Discussion

Interpretation. In a realistic reading situation, telling readers that a routine local news article was produced with AI, or entirely by AI, did not make them rate it as less credible, less professional or of lower quality, and did not make them less willing to read it. This contrasts with the penalties found in experiments based on headlines and agrees with recent findings for longer texts (Henestrosa et al., 2023; Gilardi et al., 2026; Gallegos et al., 2026). The model in Section 2.2 offers a simple account of the difference. With a full article in front of them, readers judge the text itself, so the weight on any signal about the production process is small. In addition, a byline label moves beliefs only for the minority who process it, and it often confirms what readers already assume. Each of these factors shrinks the effect of the label, and together they may explain why the effect is too small to detect.

Implications for disclosure policy. The Act exempts AI-assisted text under editorial responsibility, and publishers may see the exemption as a way to avoid a costly label. Our results suggest that, for routine news, the label itself is not costly. Neither the voluntary label for reviewed text (arm B) nor the mandatory label for AI-generated text (arm C) lowered evaluations, and the wording made no difference, although the stronger wording was much more likely to be noticed and remembered. Fear of losing credibility is therefore a weak reason to avoid disclosure. The exemption still matters for publishers, but mainly through the cost of keeping a person responsible for each text rather than through reader reactions. The labels also did not reduce demand for the article: the share of readers who would read it to the end, the margin that matters for advertising and subscription revenue (Chiou and Tucker, 2013), was the same in

all arms.

The salience trade-off. The recall results point to a tension for regulators. The purpose of disclosure is to inform readers, and a label in the byline reached the conscious awareness of only about a third of readers in the strongest version. A more prominent format would likely inform more readers, and it might also carry a larger credibility cost, because part of the reason for the small effect may be that many readers did not notice the label. Our results apply to labels of the kind a publisher would place in a byline. Testing how the prominence of a label affects both awareness and credibility is a natural next step, and it bears directly on how the Code of Practice on labelling is applied ([European Commission, 2026a](#)).

Limitations. Three limitations qualify these conclusions. First, we tested one routine, factual and neutral article; the effects may be larger for political, investigative or emotional news, or for opinion, and the design can be repeated with other stories at low cost. Second, the samples come from an online panel and are younger, more educated and more familiar with AI than national populations, although treatment effects estimated in online samples often generalize well ([Coppock et al., 2018](#)). Third, we measure stated evaluations and intentions after a single exposure in a survey, not behavior on a live site over time; a field test on a news portal would be the natural next step.

8. Conclusion

Disclosure mandates for AI-generated news are often expected to cost publishers credibility. A randomized experiment in seven countries finds that AI labels in the byline of a news article did not reduce how credible, professional or good readers found it, or whether they would read it. Most readers assumed that AI had helped produce the article even when only a journalist was named, and most did not remember which label they had seen. The estimates are close to zero and, for the two labels pooled, statistically smaller than a conventionally small effect. They do not differ significantly between the EU, where disclosure is now required, and the UK and the US, and we find no significant differences by age, education, gender or use of AI, even though these groups differ widely in their general attitudes toward AI-made news. For routine news on online portals, a disclosure mandate of the kind set by the AI Act does not appear to impose a measurable credibility cost. Replications with other kinds of news, more prominent labels, known brands and repeated exposure on live news sites would show how far this conclusion extends.

Declarations

Competing interests: Both authors are affiliated with NewsLabs Group, which develops the AI assistant named in the treatment labels. To make the results easy to verify, we fixed the primary outcome before estimation, report all collected outcomes and make the data, code and materials publicly available.

Data and materials: The anonymized data set, the analysis code, and the source code and stimuli of the survey application in all languages are publicly available, so that every table and figure can be reproduced and the experiment can be repeated independently.

Participants: Respondents were told at the start that their answers were anonymous and would be used for research only, and they were debriefed at the end as described in the notes to Table 1.

References

- Abadie, A. (2020). Statistical nonsignificance in empirical economics. *American Economic Review: Insights*, 2(2):193-208.
- Altay, S. and Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus*, 3(10):pgae403.
- Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training projects. *Journal of the American Statistical Association*, 103(484):1481-1495.
- Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727-753.
- Appelman, A. and Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, 93(1):59-79.
- Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Economic Field Experiments*, volume 1, pages 73-140. North-Holland.
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608-650.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289-300.
- Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2):889-942.
- Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809-825.
- Chiou, L. and Tucker, C. (2013). Paywalls and the demand for news. *Information Economics and Policy*, 25(2):61-69.
- Coppock, A., Leeper, T. J., and Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49):12441-12446.
- DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177-188.
- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114-126.
- Dranove, D. and Jin, G. Z. (2010). Quality disclosure and certification: Theory and practice. *Journal of Economic Literature*, 48(4):935-963.
- Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403-417.
- European Commission (2026a). Code of practice on marking and labelling of AI-generated content. Published 10 June 2026. Brussels: European Commission.
- European Commission (2026b). Commission guidelines on the implementation of the transparency obligations for certain AI systems under Article 50 of Regulation (EU) 2024/1689 (AI Act). C(2026) 5054 final, Brussels, 20 July 2026.

- European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689, 12 July 2024.
- European Parliament and Council of the European Union (2026). Regulation (EU) 2026/1744 amending regulation (EU) 2024/1689 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). Official Journal of the European Union, L 2026/1744, 24 July 2026.
- Gallegos, I. O., Shani, C., Shi, W., Bianchi, F., Gainsburg, I., Jurafsky, D., and Willer, R. (2026). Labeling messages as AI-generated does not reduce their persuasive effects. *PNAS Nexus*, 5(2):pgag008.
- Gentzkow, M. and Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2):280-316.
- Gilardi, F., Di Lorenzo, S., Ezzaini, J., Santa, B., Streiff, B., Zurfluh, E., and Hoes, E. (2026). Quality perceptions and intended engagement in response to AI-generated and AI-assisted news. *Scientific Reports*, 16:29929.
- Graefe, A., Haim, M., Haarmann, B., and Brosius, H.-B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, 19(5):595-610.
- Grossman, S. J. (1981). The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, 24(3):461-483.
- Hainmueller, J., Hangartner, D., and Yamamoto, T. (2015). Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences*, 112(8):2395-2400.
- Henestrosa, A. L., Greving, H., and Kimmerle, J. (2023). Automated journalism: The effects of AI authorship and evaluative information on the perception of a science journalism article. *Computers in Human Behavior*, 138:107445.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65-70.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge.
- Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5-86.
- Jia, H., Appelman, A., Wu, M., and Bien-Aimé, S. (2024). News bylines and perceived AI authorship: Effects on source and message credibility. *Computers in Human Behavior: Artificial Humans*, 2(2):100093.
- Jin, G. Z. and Leslie, P. (2003). The effect of information on product quality: Evidence from restaurant hygiene grade cards. *Quarterly Journal of Economics*, 118(2):409-451.
- Kirk, C. P. and Givi, J. (2025). The AI-authorship effect: Understanding authenticity, moral disgust, and consumer responses to AI-generated marketing communications. *Journal of Business Research*, 186:114984.
- Kling, J. R., Liebman, J. B., and Katz, L. F. (2007). Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83-119.
- Kolodinsky, J. and Lusk, J. L. (2018). Mandatory labels can improve attitudes toward genetically engineered food. *Science Advances*, 4(6):eaag1413.

- Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355-362.
- Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071-1102.
- Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, 7(1):295-318.
- Loewenstein, G., Sunstein, C. R., and Golman, R. (2014). Disclosure: Psychology changes everything. *Annual Review of Economics*, 6:391-419.
- Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90-103.
- Longoni, C., Fradkin, A., Cian, L., and Pennycook, G. (2022). News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 97-106. ACM.
- MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305-325.
- Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57-91.
- Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12(2):380-391.
- Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A., and Nielsen, R. K. (2024). Reuters Institute Digital News Report 2024. Report, Reuters Institute for the Study of Journalism, University of Oxford.
- Newman, N., Ross Arguedas, A., Robertson, C. T., Nielsen, R. K., and Fletcher, R. (2025). Reuters Institute Digital News Report 2025. Report, Reuters Institute for the Study of Journalism, University of Oxford.
- Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187-192.
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., and Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4):1643-1662.
- Raj, M., Berg, J. M., and Seamans, R. (2026). The artificial intelligence disclosure penalty: Humans persistently devalue AI-generated creative writing. *Journal of Experimental Psychology: General*, 155(4):896-915.
- Rosenberger, W. F. and Lachin, J. M. (2016). *Randomization in Clinical Trials: Theory and Practice*. Wiley, Hoboken, NJ, second edition.
- Scholz, F. W. and Stephens, M. A. (1987). K-sample Anderson-Darling tests. *Journal of the American Statistical Association*, 82(399):918-924.
- Simon, F. M. (2024). Artificial intelligence in the news: How AI retools, rationalizes, and reshapes journalism and the public arena. Report, Tow Center for Digital Journalism, Columbia University.
- Stantcheva, S. (2023). How to run surveys: A guide to creating your own identifying variation and revealing the invisible. *Annual Review of Economics*, 15:205-234.

- Toff, B. and Simon, F. M. (2025). “Or they could just not use it?”: The dilemma of AI disclosure for audience trust in news. *The International Journal of Press/Politics*, 30(4):881-903.
- Wittenberg, C., Epstein, Z., Péloquin-Skulski, G., Berinsky, A. J., and Rand, D. G. (2025). Labeling AI-generated media online. *PNAS Nexus*, 4(6):pgaf170.
- Young, A. (2019). Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics*, 134(2):557-598.

A. Additional tables and figures

Table A1. Covariate availability by market (percent of respondents with a valid value)

Covariate	UK	US	Croatia	Italy	Germany	France	Spain
News reading frequency (1 to 5)	100	100	99	100	100	100	100
Age (years)	100	98	100	99	99	100	99
Female	99	98	97	99	98	99	98
Tertiary education	100	98	100	99	98	99	98
Subjective social status (1 to 10)	100	98	100	99	99	100	99
AI use at work (0 to 4)	100	98	100	99	99	100	99
Student	99	97	88	99	98	99	99
Employed full-time	99	97	90	99	99	99	99
Born in country of residence	100	98	100	99	99	100	99
First language is survey language	100	98	100	99	99	100	99
Panel experience, log(1 + studies)	100	100	100	100	100	100	100
Mobile device	100	100	100	100	100	100	100
Personal income, within-country percentile	94	96	0	92	93	96	95

Notes: Share of assigned respondents with a valid value. Profile values are missing when a respondent had withdrawn consent to share a profile item, the item had expired, or the respondent preferred not to answer. Income is not part of the Croatian panel profile. In the regressions, missing values are replaced by the sample mean and an indicator for missingness is added.

Table A2. Balance of pre-treatment covariates across arms

Covariate	Mean by arm			Normalized difference		<i>p</i> -value		<i>N</i>
	A	B	C	B – A	C – A	Means	Distribution	
News reading frequency (1 to 5)	4.22	4.27	4.34	0.061	0.132	0.304	0.332	678
Age (years)	36.74	35.29	36.29	−0.136	−0.042	0.309	0.354	674
Female	0.44	0.44	0.34	0.009	−0.206	0.031	0.038	668
Tertiary education	0.60	0.62	0.65	0.047	0.099	0.590	0.588	671
Subjective social status (1 to 10)	5.52	5.30	5.65	−0.140	0.079	0.072	0.133	674
AI use at work (0 to 4)	2.08	2.21	2.32	0.098	0.183	0.150	0.201	674
Student	0.16	0.20	0.15	0.103	−0.032	0.355	0.340	661
Employed full-time	0.60	0.68	0.59	0.154	−0.024	0.117	0.134	664
Born in country of residence	0.77	0.79	0.78	0.035	0.028	0.937	0.939	674
First language is survey language	0.83	0.85	0.88	0.065	0.136	0.349	0.352	674
Panel experience, log(1 + studies)	5.18	5.20	5.10	0.012	−0.059	0.679	0.819	679
Mobile device	0.82	0.81	0.77	−0.022	−0.115	0.451	0.437	679
Personal income, within-country percentile	0.48	0.49	0.55	0.056	0.243	0.046	0.021	573

Notes: Normalized difference: difference in means divided by the pooled standard deviation. *Means:* *p*-value of a joint test that both arm coefficients are zero in a regression of the covariate on arm indicators and market fixed effects (HC2). *Distribution:* *k*-sample Anderson-Darling test of equal distributions across the three arms, by permutation. Joint tests across all covariates: multinomial logit likelihood ratio test $\chi^2(40) = 53.0$, $p = 0.08$; randomization test of the sum of the covariate-by-covariate *F* statistics, $p = 0.06$.

Table A3. Assignment, completion and balance by market

Market	Assigned				Completed	Omnibus balance test	
	Arm A	Arm B	Arm C	Total		Statistic	Randomization p
United Kingdom	33	34	34	101	100	11.2	0.587
United States	34	35	34	103	100	26.9	0.007
Croatia	24	24	24	72	71	10.9	0.558
Italy	34	34	33	101	100	19.8	0.058
Germany	34	34	33	101	100	12.6	0.463
France	34	33	34	101	100	15.2	0.295
Spain	34	33	33	100	99	10.7	0.696
All markets	227	227	225	679	670	19.5	0.058

Notes: The omnibus statistic is the sum over covariates of the F statistic for equal means across arms. Its p -value is computed by permuting arm labels within randomization blocks (1,000 permutations within markets, 2,000 for the pooled sample, which also includes market fixed effects).

Figure A1. Survey screens after the article (United Kingdom version)

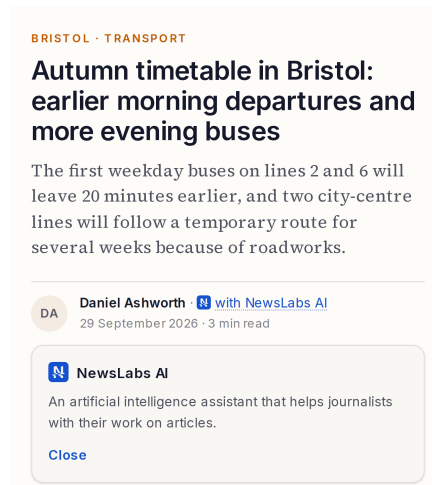
Notes: Top row, from left: news reading frequency (Q1, before the article), scene setting (Q2), impression items (Q4), reading intention (Q5). Bottom row: recall of the byline (Q6), attention check (Q7), belief about production (Q8), general trust (Q9).

Table A4. Ordered logit estimates

Outcome	B: Journalist with AI		C: Created by AI		N
	Coef.	(SE)	Coef.	(SE)	
Overall quality (1 to 7)	0.012	(0.176)	−0.011	(0.177)	675
Credibility (1 to 7)	−0.120	(0.176)	−0.053	(0.179)	675
Professional writing (1 to 7)	−0.188	(0.175)	−0.165	(0.176)	675
Reading intention (1 to 4)	0.038	(0.192)	0.063	(0.194)	675

Notes: Coefficients are log odds of a higher category. All models include market fixed effects and pre-treatment covariates.

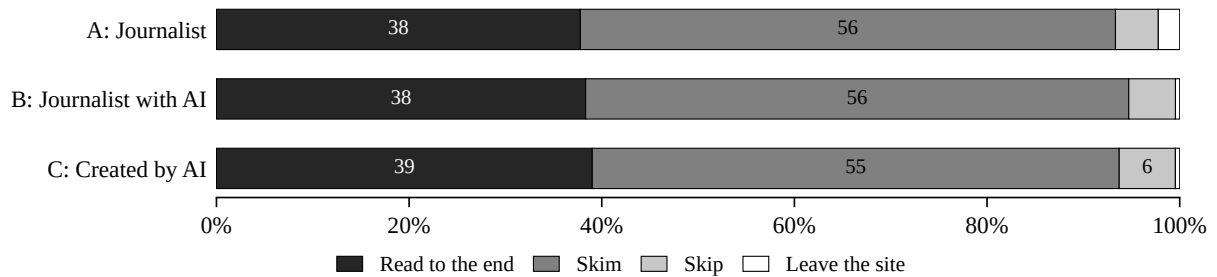
Figure A2. Explanation card in arm B



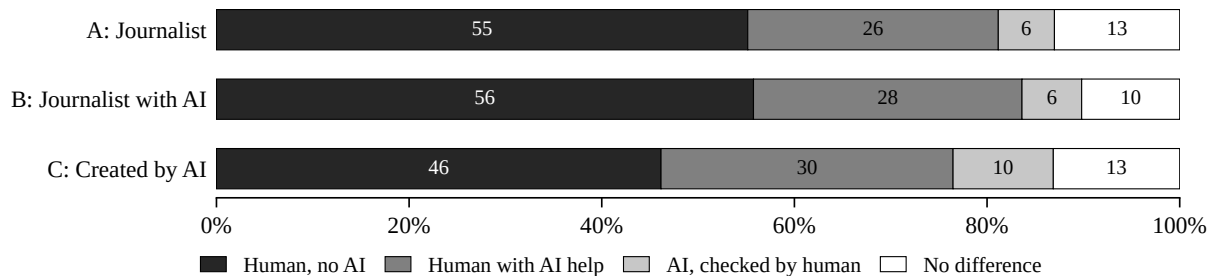
Notes: The card opened when a respondent in arm B tapped “with NewsLabs AI”. No respondent opened it.

Figure A3. Reading intention and general trust by arm

(a) Reading intention: “What would you most likely do?”



(b) General trust: “Which would you generally trust most?”



Notes: Percent of respondents in each arm. Panel (b) is measured after treatment (Q9) and introduced as a question about general views.

Table A5. Effects on general trust by mode of production (Q9, secondary outcome)

	Human without AI		AI-assisted forms		No difference	
	(1)	(2)	(3)	(4)	(5)	(6)
B: Journalist with AI	0.004 (0.047) [0.967]	0.002 (0.046)	0.025 (0.044) [0.538]	0.031 (0.044)	−0.029 (0.030) [0.347]	−0.033 (0.030)
C: Created by AI	−0.092 (0.047) [0.062]	−0.083 (0.047)	0.091 (0.045) [0.043]	0.079 (0.045)	0.001 (0.032) [0.952]	0.004 (0.031)
<i>p</i> -value, C = B	0.042	0.075	0.139	0.293	0.325	0.232
Mean in arm A	0.55	0.55	0.32	0.32	0.13	0.13
Observations	670	670	670	670	670	670
Covariates	No	Yes	No	Yes	No	Yes

Notes: Linear probability models for the option a respondent trusts most in general. *AI-assisted forms:* news written by a human with the help of AI, or produced by AI and checked by a human. All models include market fixed effects. HC2 standard errors in parentheses; randomization inference *p*-values in brackets (specifications without covariates).

Table A6. Effects on the credibility index by market

Market	<i>N</i>	Control mean (1 to 7)	B: Journalist with AI		C: Created by AI	
			Coef.	(SE)	Coef.	(SE)
United Kingdom	100	5.48	0.287	(0.229)	0.119	(0.250)
United States	102	6.15	−0.336	(0.202)	−0.214	(0.233)
Croatia	71	5.28	0.406	(0.321)	0.046	(0.344)
Italy	101	6.01	−0.211	(0.224)	−0.218	(0.275)
Germany	101	5.69	0.294	(0.219)	0.141	(0.216)
France	100	5.63	−0.409	(0.288)	−0.083	(0.235)
Spain	100	6.05	−0.053	(0.208)	−0.258	(0.227)
Pooled, random effects			−0.016	(0.119)	−0.070	(0.093)
Heterogeneity: I^2 (%); <i>p</i> -value of <i>Q</i>			44; 0.100		0; 0.809	

Notes: Differences in means within each market, in standard deviations of the pooled control arm, with HC2 standard errors. The pooled estimate uses the random-effects estimator of DerSimonian and Laird (1986). I^2 is the share of variation across markets attributable to heterogeneity; *Q* is Cochran's test of homogeneity.

B. Stimulus text

The English (United Kingdom) version of the article is reproduced below. Other markets used translations with a local city, photograph and journalist name: Columbus (United States), Split (Croatia), Verona (Italy), Münster (Germany), Poitiers (France) and Valladolid (Spain). Photographs are from Wikimedia Commons under Creative Commons licences, with the photographer credited below the photo.

BRISTOL · TRANSPORT

Autumn timetable in Bristol: earlier morning departures and more evening buses

The first weekday buses on lines 2 and 6 will leave 20 minutes earlier, and two city-centre lines will follow a temporary route for several weeks because of roadworks.

[Byline: arm A, B or C] [Date of participation] · 3 min read

[Photograph of a city bus]

The municipal transport authority in Bristol this week published the details of its autumn timetable, which takes effect next month. The changes are minor and affect five bus lines, mostly in the early morning and in the evening.

The biggest change concerns the first weekday departures on lines 2 and 6, which will leave 20 minutes earlier than now. The first bus on line 2 will depart at 05:10 instead of 05:30, and on line 6 at 05:25 instead of 05:45. According to the authority, the change was requested by passengers who start work early, particularly shift workers.

On line 3, evening buses after 21:00 will run every 12 minutes instead of every 15. Lines 5 and 9 will follow a temporary route for several weeks because of roadworks near the main station, and the affected stops will be marked with notices.

Autumn timetable: main changes

Line		Until now	New
Bus 2	first weekday departure	05:30	05:10
Bus 6	first weekday departure	05:45	05:25
Bus 3	evening interval after 21:00	15 min	12 min
Buses 5 and 9	roadworks near the main station	regular route	temporary route

Fares remain unchanged, and the new timetables will be available at stops and on the authority's website from next week. The authority also said it will continue replacing older ticket machines at stops with models that accept card payments.

Source: Press release of the municipal transport authority on the autumn timetable (PDF).

The source link could not be opened during the study; tapping it showed a note that the source is cited but cannot be opened. The debriefing at the end of the survey read: "The article was written for this study and does not report real events. Versions differed only in how the author was credited."