Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Implied Authenticity Effect? The Impact of Explicit Labels on AI-Generated Content

Fabian Pawelczyk[1], **Drew Dimmery**[2], **Pu Yan**[3]
[1]European University Institute
[2]Hertie School
[3]Peking University
fabian.pawelczyk@eui.eu, d.dimmery@hertie-school.org, puyan@pku.edu.cn

Abstract

This study investigates how labeling AI-generated content (AIGC) influences users' perceptions of authenticity in social media environments. Motivated by emerging global regulations mandating the disclosure of AI-generated media, we designed a pre-registered survey experiment to test two main questions: (1) Do AI labels reduce the perceived authenticity of AI-generated images? (2) Does exposure to labeled content affect perceived authenticity in unlabeled images (a potential spillover effect)? The pre-analysis plan and materials are available on OSF (see link below). We conducted a survey experiment with a *German* sample (N = 877) in which participants were randomly assigned to one of three groups: a control group without labels, a process-based label group ("AI-generated"), or a harm-based label group ("Misleading"). Participants viewed twelve Instagram-style posts, six of which contained AI-generated or AI-altered content. Perceived authenticity was measured by asking whether the depicted events actually occurred. Our results show that both labeling strategies significantly reduced perceived authenticity of AI-generated images, with average reductions of about 0.27 standard deviations. We also find evidence of an *implied authenticity effect*: exposure to labeled content slightly increased perceived authenticity in unlabeled images (about one-fifth the size of the direct labeling effect). Exploratory analyses indicate that internet skills and age are associated with perceived authenticity differences. By embedding labels in Instagram-style posts, our study increases ecological validity compared to earlier work focused on headlines or generic stimuli. Situated in Germany, a frontrunner in digital platform regulation, the findings provide rare non-U.S. evidence on AI labeling and contribute directly to ongoing policy debates.

Code & PAP — https://osf.io/987v6/overview

Introduction

Labeling AI-generated content (AIGC) has emerged as a major global policy concern. In September 2025, China enacted a far-reaching regulation requiring all forms of AIGC—including text, audio, video, and images—to be labeled both explicitly (e.g., via watermarks) and implicitly through metadata, with penalties for both removing labels and mislabeling human-made content as AI-generated (Heise 2025; Cyberspace Administration of China 2025). In the United States, President Biden's 2023 Executive Order 14110 promoted watermarking standards, but its repeal by the Trump II administration in 2025 has created uncertainty (The White House 2025). The European Union's AI Act mandates labeling of publicly available AI-generated media, particularly deepfakes, with fines of up to €15 million or 3% of global annual revenue (EU 2024).

Germany represents a critical case for studying these dynamics: as a frontrunner in digital regulation and a member state directly affected by the EU AI Act's transparency mandates, evidence from this context offers not only local but also globally relevant insights into how labeling will be implemented at scale (Ramirez-Ruiz and Senninger 2025). The labeling requirements of the AI Act are scheduled to enter into force in August 2026 after a Code of Practice governing implementation requirements is published in the summer. Once these requirements are in force, they are not scheduled to be reexamined until August 2028. While the Code of Practice is still in its draft stages, it can be modified or improved in response to empirical data about the effects of labeling.

In parallel, social media platforms are adjusting their own practices. Meta recently replaced its "Made with AI" label with a generic "AI Info" tag linking to further details (Meta 2024), while TikTok has introduced labeling guidelines for creators (TikTok 2023, 2025). Labeling-based approaches are one of the primary tools platforms are using to ensure that AI-generated content does not result in deleterious effects on democracy (Ahmed et al. 2025). Yet the effects of such labels on users' perceived authenticity and behavior remain unclear.

While prior work has documented how warning labels reduce belief in flagged misinformation (Clayton et al. 2020; Wittenberg et al. 2025), little is known about their indirect consequences for unlabeled content. Our study provides the first systematic experimental test of this potential *implied authenticity effect* in the context of visual AI content, thereby extending theories of soft moderation interventions into the multimodal, image-based environments that increasingly dominate social media.

This study addresses the following two research questions:

1. How does exposure to AI-generated content labels influ-

ence the perceived authenticity of unlabeled and labeled images?

2. To what extent do different label types (process-based vs. harm-based) and participant characteristics moderate these effects on perceived authenticity?

These questions, together with related work, lead us to derive three core hypotheses that guide the empirical analysis.

Related Work and Hypotheses

Research specifically examining the effects of AI labels on AI-generated content remains limited. To build a theoretical foundation, we draw on a broader literature on the role of warning labels in mitigating the effects of misinformation[1]. A substantial body of empirical work has shown that fact-checking labels reliably reduce the likelihood of believing or sharing misinformation. These effects, although moderated by content type and label implementation, appear robust across participant characteristics such as political ideology and demographic variables (Martel et al. 2024; Clayton et al. 2020; Pennycook et al. 2020; Shen, Kasra, and O'Brien 2021; Porter and Wood 2024).

Building on these findings, Wittenberg et al. (2025) argue that AI-generated content, when labeled similarly to misinformation, may also suffer reduced credibility. This aligns with early studies on AI labels in journalism. For instance, Altay and Gilardi (2024) show that headlines labeled as "AI-generated" are perceived as less credible—even when they are factually accurate—partly because users associate the label with full automation.

However, only a few studies have explored how labels affect the perceived authenticity of visual AI-generated content on social media. Among the most relevant is Wittenberg et al. (2025), which directly informs the design of our study. Wittenberg et al. examine different labeling strategies for image posts on social media and find that all labels—both harm-based and process-based—reduce belief relative to no label. Harm-based labels highlight the potential for deception or harm, whereas process-based labels describe how the content was created. Wittenberg et al. (2025) do not, however, examine any indirect effects of labels on *unlabeled* content, a primary goal of our study. These indirect effects were found to be of critical importance in Pennycook et al. (2020) in the case of misinformation interventions. In addition, Wittenberg et al. (2025) focus on U.S. samples, and it is not clear how effects in this population may generalize to the European context.

Although their results suggest that stronger harm-based labels (such as "False" or "Manipulated") may be more effective than neutral, process-based labels (e.g., "AI-generated"), the "Misleading" label—which we use as our harm-based label—performed somewhat weaker than anticipated. Nonetheless, the theoretical rationale that harm-based warnings should reduce perceived authenticity more than descriptive labels remains plausible, and we retain this distinction in our study design.

[1]Throughout this paper, we use the term "misinformation" to also encompass "disinformation," unless specified otherwise.

H1: Process-based and harm-based labels on AI-generated content reduce the perceived authenticity of labeled images compared to unlabeled images in the control group.

H2: Harm-based labels result in greater increases in the perceived authenticity of unlabeled content compared to process-based labels.

Another key concept shaping this research is the *implied truth effect*, introduced by Pennycook et al. (2020). Their findings show that while fact-checking labels reduce belief in flagged content, they may also inadvertently increase trust in unflagged—and potentially false—information. This spillover effect could undermine the broader goals of transparency policies.

This concern is particularly salient in the context of generative AI, where platforms are unlikely to label all AI-generated content. Wittenberg et al. (2025) hypothesize that this dynamic could lead to a corresponding "implied authenticity effect" in AI contexts, though empirical testing has thus far been limited. This leads us to the primary target of our study:

H3: Exposure to labeled AI-generated content increases the perceived authenticity of similar but unlabeled content.

Given these hypotheses, we design a survey experiment to test both the direct and indirect effects of AI content labeling. The next section outlines our empirical strategy in detail.

Experimental Design

Overview and Sample

In order to test our pre-registered hypotheses, we conducted a survey experiment with a *German sample* of monthly Instagram users, recruited via the platform Prolific ($N = 900$). After excluding participants who failed the attention check, the final sample consisted of $N = 877$ respondents (437 men, 424 women, and 16 identifying as other), with a mean age of 30.68 years (SD = 8.91). The sample was largely balanced across treatment groups, except for income, where statistical tests raised concerns about imbalance. The final dataset included $N_{\text{observations}} = 10{,}524$ image-level responses. In Table 1, we show an overview of the participants' demographics.

The focus on Germany is substantively and policy-relevant. Germany is a frontrunner in digital platform regulation (e.g., NetzDG), and prior work shows that domestic evidence is particularly influential in high-capacity states of the Global North and often travels transnationally (Ramirez-Ruiz and Senninger 2025). Thus, situating our study in the German context provides insights that are both locally embedded and internationally visible in policy debates. The timing is also critical, in that it allows our results to inform the creation of the Code of Practice of the EU AI Act, which provides the details for how the transparency regulations under Article 50 should be applied.

Participants were shown a sequence of fake Instagram posts, each consisting of an image and minimal accompanying text. Given that the sample consisted of Instagram

Demographic Variable	N	%
Gender		
Male	437	49.9
Female	424	48.4
Other	16	1.7
Age Groups		
18–24	234	26.7
25–34	412	47.0
35–44	159	18.1
45–54	52	5.9
55–64	16	1.8
65+	4	0.5
Education Level (Schooling)		
High	741	84.5
Medium	71	8.1
Low	20	2.3
Other	45	5.1
Highest Degree Level (Collapsed)		
Tertiary Education	526	60.0
Still in Education	183	20.9
Secondary Education	125	14.3
No Degree / Other	42	4.8
Relative Household Income		
Low (< EUR 1,969)	293	33.4
Middle (EUR 1,969–5,250)	388	44.2
High (> EUR 5,250)	140	16.0
Missing	56	6.4
Political Ideology		
Left	289	33.0
Center	561	64.0
Right	27	3.1

Table 1: Demographic Overview of Participants

users, this represents a familiar setting for the participants. The order of images was fully randomized. Each participant viewed 12 posts, six of which were AI-generated and the other six were authentic, human-made photographs sourced from news outlets' social media accounts, depicting real events. The posts primarily focus on political topics.

Participants were randomly assigned to one of three experimental conditions: a control group that received no labels, a process-based label group in which images were labeled as "AI-generated", and a harm-based label group where selected images were tagged as "Misleading." Within each label treatment group, three of the six AI-generated images were randomly selected to carry a label. All labeled images were either fully generated or visibly altered using AI tools (with one exception that we discuss in the discussion section), as verified by the DPA fact-checking team.[2] After the survey, participants were shown a debriefing page in which each stimulus image was labeled as either *Generated with AI* or *Authentic*.

The English translation of the survey and the deviations made with reference to the pre-analysis plan (PAP) can be found in the Online Appendix located in the OSF repository.

Survey Design and Implementation

To determine a reasonable target for our sample size, we conducted a power analysis. The analysis suggested that the required sample size to detect treatment effects of 15 percentage points lies between 441 and 525 (see Appendix Figure 2).

We exhausted the full budget available for participant recruitment. This resulted in a final sample of $N = 877$ participants, distributed nearly evenly across the three experimental groups (approximately 292 per group).

Participants were recruited on Prolific between April 4, 2025, and April 14, 2025. We aimed to pay participants at a rate equivalent to £9 per hour, which is considered a fair wage according to Prolific's guidelines. Based on pilot testing, the median estimated completion time was 11 minutes. After running the full study, the actual median completion time was 10 minutes and 11 seconds, demonstrating the usefulness of piloting to avoid significant under- or overpayment. Because participants completed the study faster than expected, the effective hourly wage was £9.72. The total cost of recruitment was £2,132.32.

During the experiment, we received 930 submissions, of which 53 were excluded. Eight participants timed out (did not finish within Prolific's maximum time), twelve returned their submissions (two without consent, ten who stopped voluntarily), and ten were rejected according to Prolific's policies (failing both standard attention checks or one check combined with completion under five minutes). We recruited ten replacement participants. For our main analysis, we excluded all remaining participants who failed any of our attention checks, even if they could not be formally rejected under Prolific's rules. Contrary to our pre-analysis plan, we do not have demographic information on excluded individuals and thus cannot calculate attrition rates relative to completers. Nevertheless, only ten returned submissions out of 922 total (approved + rejected + returned) amount to approximately 1.08%, which is low compared to similar studies and suggests that the landing page and survey structure were effective (Stantcheva 2023).

In addition, we implemented an attention check not aligned with Prolific's official guidelines. Following the approach of Guess and Munger (2023), we introduced a fictitious generative AI tool called *DeepMorph*. Participants who indicated having used DeepMorph were excluded from the analysis. Twelve participants failed this check, though some also failed other checks, so this figure does not represent additional exclusions. This method helps identify professional survey takers who may pass standard checks with minimal effort, thereby improving response quality.

After excluding all invalid or rejected submissions, the final sample size is *N = 877*.

Covariate Balance As pre-specified, we conducted balance checks to assess covariate balance between treatment

[2]See https://www.dpa.com/de/faktencheck for more information.

and control groups using *t*-tests for continuous variables and chi-squared tests for categorical variables. The balance table can be found in Table 4 in the Appendix.

Overall, covariate balance was good across groups. All covariates were well balanced, with no significant differences (all $p > 0.12$).

Additionally, we evaluated joint covariate balance using omnibus likelihood ratio tests based on logistic regression models, with age, income group, internet skills, AI skills, AI use, social media use, political ideology, education level, and gender as predictors.

For the process-based label group, covariates jointly predicted group assignment at a marginal level ($p = 0.101$), suggesting slight imbalance, although the corresponding McFadden's pseudo-R^2 was low (approximately 0.028), indicating minimal practical impact. For the harm-based label group, covariates were unrelated to treatment assignment ($p = 0.467$), indicating good balance.

Experimental Conditions and Stimulus Design After the consent page, all participants received instructions explaining that they would be shown a series of images and asked to indicate whether the events depicted in those images actually occurred. Participants were then randomly assigned to one of three experimental groups, with the order of images fully randomized within each group:

- **Baseline group:** Participants viewed all 12 images without any labels.
- **Process-based label group:** Participants saw a label reading "AI-generated" with the accompanying text: "This post contains media generated with artificial intelligence." *(translated from German)*
- **Harm-based label group:** Participants saw a label reading "Misleading" with the accompanying text: "This post contains media generated with artificial intelligence and may be misleading." *(translated from German)*

This design deviates from our pre-analysis plan, in which we proposed a fourth group that would receive a combined label treatment (incorporating both process-based and harm-based elements), as previously found to be highly effective in related research (Wittenberg et al. 2024). We omitted this group to maintain statistical power within our available budget. We also deviated from the PAP's suggestion to manipulate image order so that at least one labeled image would appear within the first three images shown. Due to technical limitations and time constraints, full randomization was implemented instead. This deviation, however, has the advantage of minimizing potential order effects (see the Online Appendix for all deviations).

Each participant viewed 12 images, six of which were altered or generated using AI and six were authentic, human-made photographs. The stimulus set predominantly featured political and public-affairs content, spanning elite politics, protests and geopolitical conflict, alongside a small number of non-political images (e.g., sports). Approximately half of the images depicted Germany-specific contexts, while the remainder referenced international political actors or events. Table 5 provides an overview of image topics, content types, and geographic context. In the two label treatment groups, labels were randomly assigned to three of the six AI-generated images.

To ensure internal validity and isolate the causal effect of the labels, we intentionally excluded engagement metrics (e.g., likes, comments, share counts) from the stimuli. This design choice prevents the interference of social media metrics when users judge credibility based on popularity of the content rather than the perceived authenticity, which confounds the treatment effects of the AI labels.

Label Design and Justification We chose the label *AI-generated* because Epstein et al. (2023) found experimental evidence that the term "AI-generated" tends to evoke relatively neutral associations regarding the trustworthiness of content across different cultural contexts (United States, Mexico, Brazil) and seems to be informative for users without immediately implying deception.

We selected the label *Misleading* despite previous findings suggesting that it may be "too lightweight when it comes to manipulated media," and that "members of the public may benefit from greater specificity, perhaps in the form of more decisive language (e.g., 'False' versus 'Misleading')" (Wittenberg et al. 2024). Our decision was motivated by three considerations:

First, we believe that *Misleading* strikes a realistic balance—it signals potential issues with content without being overly aggressive, making it plausible for real-world deployment by platforms.

Second, previous findings regarding the effectiveness of different label wordings were largely derived from English-language samples, primarily in the United States. It remains unclear how harm-based labels such as *Misleading* are interpreted in the German language context. By including *Misleading*, we aim to explore whether these labeling dynamics translate across languages and cultural settings.

Third, the label *Misleading* was actually used in practice, notably by Twitter (now X) during earlier initiatives to flag potentially deceptive or inaccurate information, particularly during the COVID-19 pandemic and U.S. elections (NPR 2020). Thus, using *Misleading* further increases the ecological validity of our experimental design.

In the Appendix, in Figures 3-5, we present all stimuli images and their corresponding labels. An example is shown in Figure 1. The stimulus design is inspired by Wittenberg et al. (2025), who used fake Facebook posts to study the effects of various labeling strategies. Our images were sourced via the DPA fact-checking team using Google queries such as `site:dpa-factchecking.com/ KI Bild`. All AI-generated images had been shared on social media within the 12 months preceding the publication of our PAP. To conform to Instagram's visual style, some images were cropped and reformatted accordingly. Authentic images were identified manually on Instagram and broadly matched to the topics of the AI-generated images. While the majority of stimuli are related to political events, we also included one sports image to test whether labeling effects generalize beyond political content.

To increase external validity, we closely followed the visual conventions of typical Instagram posts (as of February 2025). We blurred the profile picture and username of the account to reduce the influence of credibility cues tied to identity. In future research, it would be advisable to use factorial design experiments that systematically manipulate this element in order to explore the effects of different social media post designs (Vecchiato and Munger 2025). Additionally, future studies could benefit from approaches that leverage synthetic images and generate systematically altered versions using generative image models (see Sanderson, Tucker, and Zhong (2025)).

A key distinction from earlier designs, such as Wittenberg et al. (2025), is that we also included unlabeled authentic images. This not only allows us to identify potential spillover effects on perceived authenticity in non-AI content but also better reflects real-world user environments, in which AI-generated and authentic content are intermingled.

Figure 1: Example of AI-generated misinformation: The image above falsely shows Mike Tyson holding a Palestinian flag in the ring. The post closely follows Instagram posts style. Some differences are made: Engagement metrics (e.g., number of comments or counts of users who like the post) are not included.

One example of an AI-generated image used in the experiment is shown in Figure 1. It falsely depicts Mike Tyson holding a Palestinian flag in the boxing ring following his (real) comeback fight against Jake Paul on November 15, 2024. According to the DPA fact-checking team, several visual inconsistencies indicate that the image is not authentic: Tyson's iconic face tattoo differs in shape and intensity from verified photographs; his chest tattoo commemorating his daughter—visible in public posts since December 2023—is missing; the folds and printing of the flag appear visually inconsistent with how a real flag would hang; and the sponsor logos typically seen on the boxing ring ropes are either blurred or absent. For full verification details, see: https://dpa-factchecking.com/germany/241118-99-72329/.

This design improves ecological validity compared to earlier work that relied on textual headlines or abstract mockups. By embedding AI labels in Instagram-style posts, we more closely mimic the environments where users actually encounter generative content.

Outcome Variables The primary outcome of the survey experiment is the perceived authenticity of each post, which we measured by asking participants: "Did the event shown in the image above actually take place?"

In addition to these outcomes, we collected a range of covariates, including basic demographic variables (age, gender, education) aligned with German classification standards (ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V., Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI), and Statistisches Bundesamt 2024), a short political ideology measure (adapted from Kozyreva et al. (2021) and Yan, Schroeder, and Stier (2022)), AI and social media usage patterns (frequency of use for a predefined list of AI models and social media platforms), and indicators of internet and AI skills.

While the internet skills survey items are well-established in the literature (Deursen 2010), the AI skills scale was newly constructed for this study, drawing on existing measurement approaches (Lintner 2024; Lee and Park 2024) and adapted to better capture the breadth of user competencies across different types of AIGC.

All covariates were measured *after* the treatment exposure to ensure that the stimulus was not influenced by participants' demographic or attitudinal characteristics and to minimize potential dropout before completing the core experimental task.

Analysis

We follow the specifications outlined in our pre-analysis plan. Prior to modeling, perceived authenticity ratings are rescaled to the interval $[0, 1]$, where $0 =$ *Definitely not authentic* and $1 =$ *Definitely authentic*. This normalization facilitates interpretation of regression coefficients.

For transparency and reproducibility, the full regression code can be found in the online supplementary materials while variable definitions and constructions are provided in the Appendix (see Tables 11-13).

Average Treatment Effect

To test **H1** (Effect of labels on the perceived authenticity of labeled images) and **H2** (Difference in effect between harm-based and process-based labels), we estimate the Average Treatment Effect (ATE) using a linear regression framework. Our primary specification includes two treatment dummies and image fixed effects. Standard errors are clustered at both the participant and image levels, consistent with the other approaches and findings (Pennycook et al. 2020; Abadie et al. 2023).

$$Y_{ij} = \beta_0 + \beta_1 \cdot \text{ProcessLabel}_{ij} + \beta_2 \cdot \text{HarmLabel}_{ij} + \lambda_j + \varepsilon_{ij} \quad (1)$$

Where:

- Y_{ij} is the perceived authenticity rating of image j by participant i.
- ProcessLabel_{ij} and HarmLabel_{ij} are dummy variables equal to 1 if the corresponding label type was shown.
- λ_j denotes image fixed effects to account for variation across images.
- ε_{ij} is the error term, clustered by participant and image.

Spillover Effects

To test **H3**, we examine whether exposure to labeled images affects perceived authenticity of *unlabeled* content—that is, whether a spillover effect exists. We estimate the average effect of label exposure on perceived authenticity ratings by including the binary variable `exposed_to_label`, which equals 1 if the participant was shown at least one labeled image (regardless of label type), and 0 otherwise.

The analysis is restricted to observations where the image itself was *not* labeled (`is_labeled == 0`). All models include image fixed effects to account for image-specific differences in trustworthiness, and standard errors are clustered at both the participant and image level.

The regression specification used to estimate the spillover effect is given by:

$$Y_{ij} = \beta_0 + \beta_1 \cdot \text{ExposedToLabels}_i + \lambda_j + \varepsilon_{ij} \quad (2)$$

Where:

- Y_{ij} is the perceived authenticity rating for participant i and unlabeled image j.
- ExposedToLabels_i is a binary indicator equal to 1 if participant i was exposed to at least one labeled image.
- λ_j denotes image fixed effects.
- ε_{ij} is the error term, clustered by participant and image.

This model captures whether the presence of labeled content in a participant's experience influences their perceived authenticity in images that are not explicitly flagged.

	Unadj. Model	**Preferred Model**
Process Label	-0.094**	-0.091**
	(0.020)	(0.019)
Harm Label	-0.096**	-0.093**
	(0.016)	(0.015)
Num. of participants	877	877
Num. of observations	5,262	5,262
R^2	0.156	0.168
Controls	No	Yes
Image FEs	Yes	Yes
Clustered SEs	Yes	Yes

Table 2: Average Treatment Effects (ATE) on perceived authenticity. Entries report coefficients with clustered standard errors in parentheses. The table presents an unadjusted specification and the preferred model, which includes the full set of pre-specified participant covariates. For readability, only the focal treatment coefficients are shown. The full coefficient table is reported in Appendix Table 7. Standard errors are clustered at the participant and image level. Statistical significance: * $p < 0.05$; ** $p < 0.01$.

Results

The Direct Effect of Labeling

We find that both labeling strategies significantly decrease the perceived authenticity of AI-generated images, with reductions of approximately 0.09 points on the perceived authenticity scale. These effects are robust across model specifications and remain substantial when standardized. This provides evidence in favor of **H1**. Furthermore, higher internet skills are associated with lower perceived authenticity ratings while age is associated with higher perceived authenticity ratings, but since these covariates were not randomized, we interpret these findings as descriptive associations rather than causal effects. Our main regression results are shown in Table 2. Following the recommendations of Becker (2005), we report estimates for all terms in the regression specification in the Appendix to ensure full transparency. For completeness, Appendix Figure 6 provides a descriptive visualization of average perceived authenticity across treatment conditions and participant subgroups.

To estimate the Average Treatment Effect of labeling AI-generated images, we run two regression models, both including image fixed effects and standard errors clustered at the participants and image levels. All continuous covariates are mean-centered. The analysis is restricted to AI-generated images (since only these can be accompanied by an AI label in our experiment), resulting in 5,262 observations in the primary models that we show in Table 2. In addition, we run robustness checks with all participants including the ones that failed the attention check (AC Incl.). The results of the robustness check can be found in Table 8.

The unadjusted model in our primary models (Table 2) is a baseline specification without covariates, including only two treatment indicators: one for the process-based label ("AI-generated") and one for the harm-based label ("Misleading"). In this simple specification, we observe that the pro-

cess label reduces perceived authenticity by approximately 0.094 points (SE = 0.020, $p < 0.01$) compared to unlabeled images. Similarly, the harm label reduces perceived authenticity by about 0.096 percentage points (SE = 0.016, $p < 0.01$) on the perceived authenticity scale.

Our preferred model introduces relevant covariates and further includes interaction terms between treatment and continuous covariates to improve asymptotic precision. Following Lin (2013), regression adjustment with a full set of treatment–covariate interactions ensures that adjustment cannot worsen asymptotic precision. In practice, however, fully interacting all covariates can create problems when categorical variables contain many levels or sparse cells, leading to unstable estimates and inflated variances. To balance efficiency and robustness, we include treatment interactions only with continuous covariates. Across all our specifications, the effect sizes remain remarkably stable, with final estimates indicating around a 0.091 point reduction. Standard errors also decrease slightly (from 0.020 to 0.019), suggesting improved model precision.

Given the use of a bounded outcome scale (0–1), we standardize the perceived authenticity score to facilitate comparison with previous studies. After mean-standardizing the outcome, the labeling effects correspond to a reduction of -0.27 standard deviations for the process label (95% CI: $[-0.42, -0.13]$, $p = 0.0048$) and -0.28 standard deviations for the harm label (95% CI: $[-0.39, -0.17]$, $p = 0.0015$).

Two covariates show notable associations with perceived authenticity. First, higher internet skills are associated with a statistically significant decrease in perceived authenticity toward AI-generated images, with an estimated reduction of approximately 0.03 points per unit increase in centered internet skills ($p < 0.05$). Second, age is positively associated with perceived authenticity: each additional year of age (centered) corresponds to a 0.002 percentage point increase in perceived authenticity ($p < 0.05$).

Interestingly, the coefficient for respondents identifying as "Other Gender" is large and statistically significant. However, this result should be interpreted with caution, given the very small sample size of this group (approximately 1.7% of the sample). Small group sizes can increase the likelihood of unstable coefficients.

Prior work often suggests that age-related effects in digital trust are largely mediated by differences in internet skills (see Guess and Munger (2023) for further discussion). However, our results show that even after controlling for internet skills, age remains a significant independent predictor of perceived authenticity. Thus, the association between age and perceived authenticity is not fully explained by differences in digital literacy. This pattern points to broader age-related factors, such as generational differences in technology use, which we discuss further in the discussion section.

To formally assess whether the effects of the process-based and harm-based labels differ (**H2**), we conducted a linear hypothesis test comparing their coefficients. The test fails to reject the null hypothesis that the effects are equal ($F(1, 5244) = 0.004$, $p = 0.95$), suggesting that the two types of labels exert statistically indistinguishable impacts on perceived authenticity.

Do Label Effects Vary Across Subgroups? To examine heterogeneous effects, we conduct a subgroup analysis by estimating treatment effects with our preferred model specification separately for each dimension of the categorical moderators. For continuous moderators, we split the sample into two groups based on the median (i.e., values below or equal to the median versus values above the median) and again estimate treatment effects accordingly. The results of this analysis are presented in Figure 7. In total, this corresponds to 24 separate model estimations. Four subgroups exhibit particularly wide confidence intervals due to small sample sizes: respondents with missing income information, respondents identifying with right-wing political ideology, those reporting no or other degrees, and those identifying with "other" gender. For reference on sample composition, see Table 4. Overall, the subgroup results reveal no evidence of systematic heterogeneity. The only subgroup difference that is nominally significant prior to adjustment concerns the harm-label effect among respondents with "No or Other" education. To account for multiple subgroup comparisons, we test subgroup deviations from the overall ATE and apply false-discovery-rate control using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995). This difference does not survive false-discovery correction, is based on a small residual category, and is not mirrored across other education levels or label types. We therefore interpret it as a chance finding rather than evidence of substantive heterogeneity and report the results of the false-discovery correction in Table 15.

Spillover Effects on Unlabeled Content

We find evidence that exposure to labels on AI-generated content significantly increases perceived authenticity in *unlabeled* images. This provides support for **H3**. Across model specifications, the spillover effect is positive, statistically significant, and robust to the inclusion of controls and interactions. Standardized results suggest a modest but meaningful increase in perceived authenticity. Main regression results can be found in Table 3 and the table with the same specifications reporting all covariates can be found in Table 9. Additionally, we display robustness checks in Table 10.

As specified in the pre-analysis plan, the primary estimand of this study is the spillover effect, which examines whether exposure to labeling influences perceived authenticity in subsequently encountered unlabeled content. To estimate this effect, we restrict the dataset to ratings of unlabeled images and define the treatment as prior exposure to a labeled image. This is reflected in a dummy variable that equals one for all observations after a participant saw at least one label. The final analytic sample includes 8,773 observations across the main models.

Following the strategy used for the ATE model, all continuous covariates are mean-centered, image fixed effects are included, and standard errors are clustered at the image and participant level. We add covariates and interaction terms to test robustness and improve asymptotic precision.

In the baseline model without covariates, exposure to a label increases perceived authenticity of unlabeled images

	Unadj. Model	Preferred Model
Label Exposure	0.018*	0.019*
	(0.007)	(0.006)
Num. of participants	877	877
Num. of observations	8,773	8,773
R^2	0.366	0.371
Controls	No	Yes
Image FEs	Yes	Yes
Clustered SEs	Yes	Yes

Table 3: Spillover effects on perceived authenticity in unlabeled images. Entries report coefficients with clustered standard errors in parentheses. Standard errors are clustered at the participant and image level. The preferred model includes the full set of pre-specified participant covariates. The full coefficient table and the models including interaction terms to explore potential heterogeneity are reported in Appendix Table 9. Statistical significance: * $p < 0.05$; ** $p < 0.01$.

by approximately 0.018 percentage points (SE = 0.007, $p < 0.05$). In the full specification with all covariates (Preferred Model) and relevant interactions, the estimated effect size remains stable at 0.019 percentage points (SE = 0.006, $p < 0.05$).

To facilitate comparison with prior literature, we standardize the perceived authenticity score. After standardization, the exposure effect corresponds to an increase of 0.057 standard deviations (95% CI: $[0.014, 0.100]$, $p = 0.013$).

The positive spillover effect suggests that labeling AI-generated content can have broader impacts beyond labeled posts, enhancing perceived authenticity even in unlabeled content. This finding supports **H3** and points to a potential unintended consequence of labeling policies, whereby cautionary labels increase overall perceived authenticity toward content within the same platform environment.

Do Spillover Effects Vary Across Subgroups? To examine heterogeneous spillover effects, we build on the approach described in the previous subgroup analysis section. Specifically, we estimate spillover effects with our preferred model specification separately for each dimension of the covariates. The results are presented in Figure 8. Similar to the main spillover effect, we do not find evidence of heterogeneity: the coefficient of 0.019 lies within the 95% confidence interval of each subgroup model.

What Drives the Spillover Effect? After estimating the main models, a critical question remains: what drives the spillover effect? Specifically, does exposure to labels affect perceived authenticity in AI-generated images differently than in authentic images? One possibility is that participants increase their perceived authenticity in authentic (unlabeled) content after seeing labels. Another is that participants extend more perceived authenticity to unlabeled AI-generated images, reasoning that AI content would be flagged if present.

To examine this, we also run an *AI Interaction Model* (see Table 9 in the Appendix), which includes an interaction between label exposure and AI-generated images. The coefficient on this interaction is -0.012, suggesting that spillover effects may be weaker for AI-generated content than for authentic content, although the result is not statistically significant. To probe further, we estimated separate models by content type. For authentic images, the spillover effect is positive and significant ($0.024, p < 0.05$), whereas for AI-generated images the effect is smaller and not significant (0.011).

Taken together, these results provide only weak evidence that the spillover effect is primarily driven by increased perceived authenticity in authentic images. Importantly, the study is underpowered to draw strong conclusions, and future research should examine this question with larger samples.

Discussion

Theoretical Mechanisms

Our findings compared process-based (AI-generated) and harm-based (misleading) labels and demonstrated indistinguishable reductions in perceived authenticity. Theoretically, this suggests that labels, regardless of their indication of technological details, function as part of heuristic processing. Users treat labels as generic caution cues without processing the semantic differences (Clayton et al. 2020). The mere presence of the label acts as a prompting mechanism that leads to skepticism and caution. However, we hypothesize that divergent effects of AI labels may appear in dimensions not measured in the current study. First, the "AI-generated" label explicitly signals technical origin (provenance) of the content, which may influence users' judgments of creativity rather than authenticity. Second, harm-based labels may carry a strong normative signal. Users may experience higher social reputational costs when sharing media framed as potentially harmful. While current research suggests convergent effects on perceived authenticity, future research can therefore test whether AI labels impact source attribution (provenance) or sharing intentions (social reputational costs). For instance Gallegos et al. (2025) show that AI-generated political messages do not necessarily reduce their persuasive impact, even when participants correctly recognize the content as AI-generated.

Contextualizing Effect Sizes

When comparing our findings to prior research, we observe that the direction of effects clearly aligns with the results of Wittenberg et al. (2025). Specifically, both studies find that labels reduce individuals' perceived authenticity in the events depicted in social media posts, supporting the general effectiveness of labeling strategies.

Interestingly, Wittenberg et al. report a negative effect size of -0.14 standard deviations (95% CI: $[-0.24, -0.03]$) when using the "AI-generated" label. In our study, we find similar, though slightly larger, effects — indicating a small to medium-sized reduction in perceived authenticity. Notably, Wittenberg et al. also tested stronger harm-based labels such as "Manipulated," which yielded a larger negative

effect size of -0.27 SD $[-0.37, -0.16]$. In light of these comparisons, our findings appear realistic and substantively meaningful.

It remains an open question which design features explain differences in effect sizes across studies for the same labeling terms. One notable distinction is that our design included authentic (non-AI) images alongside AI-generated content, whereas Wittenberg et al.'s design focused exclusively on manipulated content. Whether the inclusion of authentic images attenuates or amplifies labeling effects is an interesting avenue for future research.

Another explanation lies in different time contexts (December 2023 vs. April 2025). As public knowledge of AI develops rapidly, findings in this area may not remain stable over time. This reflects the broader challenge of *temporal validity* (Munger 2023), which is a particular issue for fast-moving domains such as developments in AI and social media.

In addition, the national context may play an important role. Our study relies on a German sample, and it remains unclear how cultural environments influence labeling effects. Recent surveys indicate cross-national variation in trust and attitudes towards AI (e.g., Gillespie et al. (2023)). This highlights the need for coordinated efforts to replicate similar experiments across different countries and time periods.

Regarding the implied authenticity effect, the best comparison for our results is the work by Pennycook et al. (2020), even though their study focused on headlines and fact-checking labels rather than visual content. In their first experiment, Pennycook et al. find an implied truth effect (measured on a [0,1] scale after rescaling, same approach as our study) of approximately 0.0112 (95% CI: [0.0047, 0.0177], $p < .001$), compared to a warning effect of -0.0324 (95% CI: $[-0.0129, -0.0519]$). In their second experiment, both the implied truth and warning effects grow in magnitude (due to larger, more salient labels), but the ratio between them remains similar — the implied truth effect is roughly one-third the size of the warning effect.

In contrast, in our study, the spillover effect (implied authenticity effect) is about one-fifth (0.019) of the direct labeling effect (95% CI: [0.006, 0.031], $p = .013$). While Pennycook et al. evaluate the implied truth effect primarily in relation to the size of the warning effect — a strategy that makes sense when weighing cost-benefit considerations for label implementations — we argue that this relational evaluation needs further justification.

Specifically, comparing the relative size of the warning and spillover effects informs platform design but does not fully capture the potential harm that increased perceived authenticity in unlabeled content can cause. Since labeling coverage is unlikely to be complete in real-world environments, even small absolute increases in perceived authenticity toward unlabeled misinformation could scale into meaningful societal risks. Thus, policy discussions should consider not only the relative but also the absolute magnitude and contextual risk of spillover effects.

A potential extension of this study is an addition that Pennycook et al. (2020) introduced, which is a group in one treatment which adds "Verified" labels in addition to the warning labels. A similar label in our case could be, for example "Human-made". In their experiment they find that the verified label indeed, in line with expectation, soaks up the implied truth effect and so also delivers a potential solution to counter the implied authenticity effect.

Limitations and Future Work

Age, Internet Skills and Sample Bias One important issue with interpreting our effect sizes is that online samples tend to exclude many older, low-skilled users (Guess and Munger 2023). This is particularly critical because exploratory analyses suggest that age is associated with variation in perceived authenticity for both labeled and unlabeled content. Although our internet skills index captures latent digital skills well, as indicated by a high Cronbach's alpha score, age still emerges as a substantial driver in reducing perceived authenticity ratings.

Despite our efforts to enhance external validity by filtering for Instagram users only, sample bias remains a concern. As shown in Table 1 and Table 14, the 55–64 age group comprises only 1.8% of the sample versus 9% in the population, and those aged 65 and older account for just 0.5% compared to 4% (NapoleonCat 2025). The underrepresentation is striking—especially because the few older participants we do capture likely have higher internet literacy compared to the general older population. Furthermore, the sample's mean internet skills score of 4.30 (on a 5-point scale) indicates a highly digitally literate population overall. Future work should therefore aim to obtain samples that more closely reflect the actual distribution of key user covariates.

This sampling bias is problematic because it limits our ability to understand the effects of labels among the most vulnerable user groups. We recommend that future research diversify recruitment strategies—such as recruiting via Facebook ads or conducting experiments in local libraries' computer courses—as suggested and done by Guess and Munger (2023).

The Blurred Boundaries of Image Authenticity Another important avenue for future investigation is whether the spillover effect interacts with the type of image shown (authentic vs. AI-generated). It makes a substantial difference whether increased perceived authenticity primarily concerns authentic images or AI-generated content. Our preliminary analyses suggest that such an interaction may exist, but we currently lack sufficient statistical power to draw firm conclusions. Nonetheless, this question is critical for future work.

In addition, it is essential to examine how treatment effects vary across individual images. Real-world images are complex, and the distinction between AI-generated, digitally altered, and authentic content is often blurred. For example, one image in our sample showing Greta Thunberg with a "Fuck Israel" sign was actually digitally altered from an original sign reading "Free Palestine." While we categorized this as an altered image during debriefing, it could also represent a case of traditional photo editing rather than pure AI generation. Similarly, the authentic image from the Georgian

protests (depicting a person in front of flames while wearing a gas mask) may also involve substantial photo editing or enhancement with advanced techniques.

Such examples highlight that labeling is not always straightforward: for these posts, the label "AI-generated" might be misleading, while the label "Misleading" could be more appropriate. These complexities emphasize the need for more nuanced label design and a deeper understanding of content authenticity classifications.

Advancing Experimental Designs Our study relies on perceived authenticity as the primary outcome. However, the impact of labeling may extend beyond this. Future research could distinguish between *authenticity* (is the content real?) and *perceived AI provenance* (is the media synthetic or authentic?). Users might correctly identify content as AI-generated yet still perceive the event as factually accurate. Additionally, we did not measure common social media engagement behaviors such as likes, shares, or comments. Labeling might reduce perceived authenticity in content while having little or even no effect on its social media sharing, or conversely, users might share labeled content despite knowing it is synthetic. We acknowledge that capturing these user behaviors is crucial for evaluating the full social impact of AI transparency efforts. Accordingly, we plan to incorporate social media engagement cues in future experiments on labeling. These features likely interact with label effects in real-world settings.

Concretely, such a design could employ a visual conjoint framework (Vecchiato and Munger 2025) that crosses label presence (label vs. no label) with experimentally varied popularity cues (such as likes and shares) and identity signals (e.g., blurred vs. visible profile pictures and usernames). This approach would allow systematic estimation of how social endorsement and source cues moderate the effects of AI transparency labels on perceived authenticity.

Finally, self-reported measures cannot fully capture the cognitive processing of labels in real time. To address this limitation, we conducted two follow-up studies that used eye-tracking technology to map attention distribution mechanisms during naturalistic social media scrolling (Chen et al. 2026). This approach aims to determine whether the efficacy of labels stems from attentional blindness (i.e., users fail to notice AI labels) or heuristic processing (i.e., users noticing the label but fail to process the semantic details). Integrating physiological data with behavioral outcomes will be powerful for evaluating the full impact of AI transparency efforts.

Beyond the avenues already discussed, we highlight several directions for future research on labeling AI-generated social media content.

First, related work on label design (Gamage et al. 2025), label source (Horne 2025), and explanatory mechanisms (Epstein et al. 2022) should be extended from text-based misinformation to image-based settings and should be evaluated accordingly.

Second, building on Pennycook et al. (2020), researchers could test authenticity labels (e.g., "Human-made" or "Verified") alongside AI-generated labels. Such labels might mitigate unintended trust boosts in unlabeled content by affirming authenticity.

Third, external validity can be enhanced by calibrating the share of AI-generated versus authentic content to real-world feeds and aligning topic selection (e.g., political posts) with platform-specific usage patterns.

Finally, more naturalistic user experiences could be incorporated, such as enabling scrolling behavior or using eye-tracking to study attention to labels. These additions may clarify how users cognitively process labeled content.

Overall, future work should balance experimental control with ecological validity to better capture how labeling interventions shape perceived authenticity in real-world environments.

Conclusion

Ensuring the authenticity of social media content is a critical step toward transparency and accountability. The unprecedented viral spread of AI-generated content poses new challenges to content authenticity. Our study stems from existing literature on information disclosure and the platforms' role in content moderation. Findings from this study enrich the discussion about policy recommendations for platform designers, policymakers, and regulators.

Both process-based labels ("AI-generated") and harm-based labels ("Misleading") significantly reduced the perceived authenticity of AI-generated content in our study. This suggests that labeling AI-generated media can be an effective tool to alert users and reduce the believability of misleading or synthetic content.

In addition, our results highlight the importance of maximizing labeling coverage. Although labels successfully decreased perceived authenticity, we also observed a small but measurable spillover effect: exposure to labeled images slightly increased perceived authenticity in unlabeled content. Since it is unlikely that labeling can achieve full coverage in real-world environments, efforts should focus on labeling as much AI-generated content as possible to minimize potential spillover risks.

The question of potential coverage is an important contribution to the discussion over the EU AI Act's Code of Practice. Given an estimate of the coverage of a labeling scheme, the results from this study can be used to help determine a cost-benefit ratio of labeling (at least in Germany): the benefits accrue from reduced belief in the authenticity of labeled inauthentic media, but this must be balanced against the cost of increased belief in the authenticity of unlabeled media, which may itself be inauthentic.

Moreover, the choice of label wording requires careful consideration. Although stronger harm-based labels like "False" or "Manipulated" might achieve greater reductions in perceived authenticity, they could also trigger political backlash or reduce overall trust in platform interventions. "Misleading," despite being a relatively mild label, may strike a pragmatic balance between warning users without alienating them. Decision-makers should weigh these trade-offs carefully when designing labeling frameworks.

Finally, it is important to account for the specific vulnerabilities of digitally less literate populations. Older and less

digitally skilled users were underrepresented in our sample and are often excluded from online panel studies, yet they may be particularly susceptible to misleading AI content. Platforms and regulators should ensure that labeling strategies are comprehensible and effective for diverse populations, not just for younger or more internet-savvy users.

In conclusion, while AI labels are not a complete solution to the challenges posed by generative AI and misinformation, our results suggest that they are a useful and viable tool, so long as their benefits are appropriately examined with respect to their costs.

Looking forward, our study should be seen as a contribution to the broader research agenda on labeling design and effects. By documenting direct and spillover effects in a German Instagram context, we add empirical evidence that complements existing work on soft moderation interventions. In this way, we help set the stage for comparative experiments across platforms, countries, and demographic groups.

Acknowledgments

We appreciate the contributions of all participants in our study. We are grateful for the constructive feedback from anonymous reviewers. We would also like to thank the Hertie School in Berlin, National Natural Science Foundation of China (No.72304017), Beijing Municipal Social Science Foundation (No.22XCC013) and China Association for Science and Technology (No.2023QNRC001) for funding support. Drew Dimmery was previously employed at Meta. We acknowledge the use of OpenAI's ChatGPT, which was employed to check grammar, improve clarity, and suggest alternative formulations of text. In addition, ChatGPT was used to assist with code generation, table formatting, and creating BibTeX entries for websites. All final content and interpretations are the responsibility of the authors.

References

Abadie, A.; Athey, S.; Imbens, G. W.; and Wooldridge, J. M. 2023. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1): 1–35.

ADM Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V.; Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI); and Statistisches Bundesamt. 2024. *Demographische Standards: Ausgabe 2024*, volume 31 of *GESIS-Schriftenreihe*. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften, 7., überarbeitete und erweiterte auflage edition. A joint recommendation of ADM, ASI, and Statistisches Bundesamt.

Ahmed, A.; Doyle, O.; Harris, D. E.; and Norden, L. 2025. Tech companies pledged to protect elections from AI — here's how they did. https://www.brennancenter.org/our-work/research-reports/tech-companies-pledged-protect-elections-ai-heres-how-they-did. Brennan Center for Justice. Accessed: 2025-01-08.

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.

Becker, T. E. 2005. Potential Problems in the Statistical Control of Variables in Organizational Research: A Qualitative Analysis With Recommendations. *Organizational Research Methods*, 8(3): 274–289.

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 (Methodological)*, 57(1): 289–300.

Chen, N.; Lai, Z.; Liu, Y.; Li, J.; Wang, R.; and Yan, P. 2026. See, trust, and interact: How AI disclosure shapes high school students' trust. *Information Research: An International Electronic Journal*, 31(iConf): 1099–1145.

Clayton, K.; Blair, S.; Busam, J. A.; Forstner, S.; Glance, J.; Green, G.; Kawata, A.; Kovvuri, A.; Martin, J.; Morgan, E.; et al. 2020. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4): 1073–1095.

Cyberspace Administration of China. 2025. Answers to Reporters' Questions on the Measures for Identifying Synthetic Content Generated by Artificial Intelligence. https://www.cac.gov.cn/2025-03/14/c_1743654685896173.htm. Accessed: 2025-12-31.

Deursen, A. v. 2010. Measuring internet skills. *International Journal of Human-Computer Interaction*, 26(10): 891–916.

Epstein, Z.; Fang, M. C.; Arechar, A. A.; and Rand, D. G. 2023. What label should be applied to content produced by generative AI? https://doi.org/10.31234/osf.io/v4mfz. PsyArXiv:v4mfz.

Epstein, Z.; Foppiani, N.; Hilgard, S.; Sharma, S.; Glassman, E.; and Rand, D. 2022. Do explanations increase the effectiveness of AI-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, 183–193.

EU. 2024. Art. 50 KI-VO - Transparenzpflichten für Anbieter und Betreiber bestimmter KI-Systeme.

FORCE11. 2020. The FAIR Data Principles. Accessed: 2026-03-22.

Gallegos, I. O.; Shani, C.; Shi, W.; Bianchi, F.; Gainsburg, I.; Jurafsky, D.; and Willer, R. 2025. Labeling messages as AI-generated does not reduce their persuasive effects. *arXiv preprint arXiv:2504.09865*.

Gamage, D.; Sewwandi, D.; Zhang, M.; and Bandara, A. K. 2025. Labeling Synthetic Content: User Perceptions of Label Designs for AI-Generated Content on Social Media. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–29.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gillespie, N.; Lockey, S.; Curtis, C.; Pool, J.; and Akbari, A. 2023. Trust in artificial intelligence: A global study. The University of Queensland and KPMG Australia.

Guess, A. M.; and Munger, K. 2023. Digital literacy and online political behavior. *Political Science Research and Methods*, 11(1): 110–128.

Heise. 2025. AI content must be labeled in China from September. https://www.heise.de/en/news/AI-content-must-be-labeled-in-China-from-September-10324797.html. Accessed: 2026-01-07.

Horne, B. D. 2025. Does the Source of a Warning Matter? Examining the Effectiveness of Veracity Warning Labels Across Warners. In *Proceedings of the International AAAI Conference on Web and Social Media*, 823–836.

Kozyreva, A.; Lorenz-Spreen, P.; Hertwig, R.; Lewandowsky, S.; and Herzog, S. M. 2021. Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications*, 8(1): 1–11.

Lee, S.; and Park, G. 2024. Development and validation of ChatGPT literacy scale. *Current Psychology*, 43(21): 18992–19004.

Lin, W. 2013. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1): 295–318.

Lintner, T. 2024. A systematic review of AI literacy scales. *npj Science of Learning*, 9(1): 50.

Martel, C.; Allen, J.; Pennycook, G.; and Rand, D. G. 2024. Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, 19(2): 477–488.

Meta. 2024. Our Approach to Labeling AI-Generated Content and Manipulated Media. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/. Accessed: 2026-01-07.

Munger, K. 2023. Temporal validity as meta-science. *Research & Politics*, 10(3).

NapoleonCat. 2025. Instagram users in Germany – April 2025. https://napoleoncat.com/stats/instagram-users-in-germany/2025/04/. Accessed: 2026-01-07.

NPR. 2020. Twitter Expands Warning Labels To Slow Spread Of Election Misinformation. https://www.npr.org/2020/10/09/922028482/twitter-expands-warning-labels-to-slow-spread-of-election-misinformation. Accessed: 2026-01-07.

Pennycook, G.; Bear, A.; Collins, E. T.; and Rand, D. G. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11): 4944–4957.

Porter, E.; and Wood, T. J. 2024. Factual corrections: Concerns and current evidence. *Current Opinion in Psychology*, 55: 101715.

Ramirez-Ruiz, S.; and Senninger, R. 2025. Policy documents across 185 countries predominantly rely on evidence from the Global North. https://osf.io/preprints/osf/w8q3y_v1. Accessed: 2026-01-08, OSF Preprints:w8q3y_v1.

Sanderson, Z.; Tucker, J. A.; and Zhong, W. 2025. It Works When It Works: Measuring the Direct and Indirect Effects of AI Labels on Political Images. https://osf.io/preprints/osf/nf785_v1. Accessed: 2026-01-14, OSF Preprints:nf785_v1.

Shen, C.; Kasra, M.; and O'Brien, J. 2021. This photograph has been altered: Testing the effectiveness of image forensic labeling on news image credibility. *Harvard Kennedy School Misinformation Review*.

Stantcheva, S. 2023. How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible. *Annual Review of Economics*, 15(1): 205–234.

The White House. 2025. Initial Rescissions of Harmful Executive Orders and Actions. https://www.whitehouse.gov/presidential-actions/2025/01/initial-rescissions-of-harmful-executive-orders-and-actions/. Accessed: 2026-01-07.

TikTok. 2023. New labels for disclosing AI-generated content. https://newsroom.tiktok.com/new-labels-for-disclosing-ai-generated-content?lang=en. Accessed: 2026-01-07.

TikTok. 2025. Über KI-generierte Inhalte. https://support.tiktok.com/de/using-tiktok/creating-videos/ai-generated-content. Accessed: 2026-01-07.

Vecchiato, A.; and Munger, K. 2025. Introducing the Visual Conjoint, with an Application to Candidate Evaluation on Social Media. *Journal of Experimental Political Science*, 12(1): 57–71.

Wittenberg, C.; Epstein, Z.; Berinsky, A. J.; and Rand, D. G. 2024. Labeling AI-Generated Content: Promises, Perils, and Future Directions. https://doi.org/10.21428/e4baedd9.0319e3a6. An MIT Exploration of Generative AI. Accessed: 2026-01-08.

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.

Yan, P.; Schroeder, R.; and Stier, S. 2022. Is there a link between climate change scepticism and populism? An analysis of web tracking and survey data from Europe and the US. *Information, Communication & Society*, 25(10): 1400–1439.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, our study investigates the effects of AI-content labeling on perceived authenticity. It uses survey data from consenting participants without violating privacy norms.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we note that our German Prolific sample skews young and digitally skilled, which may limit generalizability.

(e) Did you describe the limitations of your work? Yes, for instance, limitations are described regarding generalizability, sample composition, and underpowered exploratory analyses

(f) Did you discuss any potential negative societal impacts of your work? Yes, we note that labeling strategies could inadvertently increase trust in unlabeled content (spillover effect).

(g) Did you discuss any potential misuse of your work? Yes, but indirect. We acknowledge that platform designers might misapply weaker labels in ways that reduce overall trust.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, the study was preregistered, data is anonymized, and materials/code will be shared via OSF.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA, our study does not present formal theoretical results.

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, we discuss how implied authenticity builds on and complements the implied truth effect literature.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, we discuss possible alternative explanations such as sample demographics.

(e) Did you address potential biases or limitations in your theoretical framework? Yes, we note limits to extending labeling theories from text to multimodal images.

(f) Have you related your theoretical results to the existing literature in social science? Yes.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, see Discussion section.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? NA

(b) Did you mention the license of the assets? NA

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, we include the survey translation in the online appendix and we added the Instagram-styled posts into the appendix.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, we followed the ethical guidelines by the home institution; risks were identified as minimal.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, participants were compensated above "fair wage" according to Prolific's guidelines (£9.72/hour). Total costs are reported in the Survey Design section.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, Prolific IDs were anonymized, data is stored securely, and will be shared upon request.

Appendix

Covariates and Transformations

Gender The sample was nearly gender-balanced, with 49.9% of participants identifying as male, 48.4% as female,

and 1.7% identified as non-binary, selected another gender category, or preferred not to disclose their gender. For the analysis, we collapsed participants into three groups: Male, Female, and Other.

Age Participants ranged in age from 18 to 72 years, with a mean age of 30.68 (standard deviation 8.91) and a median age of 29. The majority of respondents (47%) were between 25 and 34 years old, followed by 26.7% aged 18–24, and 18.1% aged 35–44. Older age groups were less represented, with only 0.5% of participants aged 65 or older. For the statistical models, age was mean-centered to improve interpretability and model stability.

Schooling Participants were generally highly educated. A total of 84.5% reported having completed a high-level school qualification (e.g., Abitur or Fachhochschulreife), 8.1% were categorized as having medium-level education (e.g., Realschule or DDR 10th-grade equivalent), and 2.3% fell into the low education category. An additional 5.1% provided responses classified as "Other," including ongoing school attendance or non-standard qualifications. For the analysis, we collapsed schooling levels into four categories: Low, Medium, High, and Other, based on participants' highest general school qualification.

Highest Educational Degree The highest degree attained further reflects the academic skew of the sample. Over half of all participants (56.4%) held a tertiary degree, with 30.3% reporting a Bachelor's degree (classified as "Tertiary – Medium") and 26.1% holding an advanced academic qualification such as a Master's degree, Staatsexamen, or PhD (classified as "Tertiary – High"). An additional 3.7% completed applied tertiary programs such as a Meister or Technikerschule. A significant portion of the sample (20.9%) was still enrolled in education, while 14.2% reported vocational training at the secondary level or no formal professional qualification. Only 0.5% selected "Other."

For analysis purposes, we constructed a collapsed measure of educational attainment with four categories: participants were grouped as "Tertiary" if they had completed a Bachelor's degree, an advanced degree (Master's, Diplom, Magister, Staatsexamen, or PhD), or an applied tertiary program. Those who completed vocational training at the secondary level, either through a dual-system apprenticeship or a school-based vocational qualification, were grouped as "Secondary." Participants who were currently enrolled in an educational program were classified as "Still in Education," and those without a formal vocational or professional qualification, or who selected "Other," were grouped under "No or Other Degree."

Approximately 60% of participants fell into the Tertiary group, reflecting the relatively high education level of the sample. Around 14% were categorized under Secondary education, roughly 21% reported that they were still enrolled in education, and only a small fraction were classified as having no or another type of degree.

Household Income Participants reported their monthly net household income by selecting one of 23 predefined brackets, ranging from "under €500" to "€25,000 or more." Each bracket was assigned a corresponding midpoint value in euros for analysis purposes. Based on these midpoints, the sample median income was calculated.

Using the sample median (€2625) as a reference point, participants were classified into three relative income groups: Low income (household income below 75% of the median), Middle income (household income between 75% and 200% of the median), and High income (household income above 200% of the median). Participants who preferred not to disclose their income were categorized as missing. This relative income grouping was used as a categorical covariate in subsequent analyses.

Internet Skills Participants' digital literacy was measured using nine items, each rated on a 5-point Likert scale ranging from "Not at all familiar" (1) to "Very familiar" (5). The items covered four dimensions of internet skills: operational, informational, strategic, and participatory.

For analysis, we constructed a single global Internet Skills Index by averaging the nine items after numeric recoding. The internal consistency of the scale was high (Cronbach's $\alpha = 0.87$), indicating strong reliability. Missing values were handled listwise.

The final Internet Skills Index ranges from 1 to 5, with a sample mean of 4.30 (SD = 0.57), a minimum observed value of 1.56, and a maximum of 5. For regression analysis, the index was mean-centered to facilitate interpretation, and is referred to as Internet Skills (c) throughout the results.

AI Skills Participants' self-assessed AI literacy was measured using eleven items, each rated on a 5-point Likert scale ranging from "Not at all familiar" (1) to "Very familiar" (5). The items captured different dimensions of AI use, including technical tasks (e.g., training or fine-tuning models), communication with AI tools, and creative applications (e.g., generating text or images).

To construct a unified AI Skills Index, all items were numerically recoded and averaged into a single continuous score. The internal consistency of this index was excellent, with a Cronbach's $\alpha = 0.91$, indicating high reliability. Two participants indicated in a preceding filter question that they had no familiarity with AI tools; for these cases, a value of 1 was assigned on both the AI Skills Index.

The resulting AI Skills Index ranges from 1 to 5, with a sample mean of 3.39 (SD = 0.74), a minimum observed value of 1, and a maximum of 5. For regression analyses, the AI Skills Index was mean-centered to facilitate interpretation.

AI Use Participants were asked how frequently they used ten widely available generative AI platforms and tools, including ChatGPT, MidJourney, Gemini, and Suno. Response options ranged from "Never" (0) to "Multiple times per day" (4). A composite AI Use Index was constructed by averaging responses across all platforms, resulting in a continuous measure of AI engagement.

Internal consistency is not applicable for this index due to the platform-based structure of the items. Missing values were handled consistently with the AI Skills Index: partic-

ipants who indicated no familiarity with AI tools were assigned a value of 0.

The resulting AI Use Index ranges from 0 to 4, with a sample mean of 0.55 (SD = 0.40) and observed values between 0 and 2.9. For regression analyses, the AI Use Index was mean-centered to facilitate interpretation.

Social Media Use To capture broader patterns of digital behavior, participants were asked how frequently they used 14 popular social media platforms, including X (formerly Twitter), TikTok, YouTube, and others. Usage for each platform was recorded on a 5-point scale ranging from "Never" (0) to "Multiple times per day" (4).

A Social Media Use Index was constructed by averaging responses across all platforms, producing a continuous measure of general social media engagement. The resulting index ranges from 0 to 4, with observed values between 0.07 and 3.00 (Mean = 1.10, SD = 0.40).

For regression analyses, the Social Media Use Index was mean-centered.

Political Ideology Participants were asked to position themselves on a standard 7-point left–right political spectrum, where 1 indicated the political left, 4 the center, and 7 the political right. The question prompt explained that such labels are often used to describe political parties, ideologies, and political leaders.

For our models, we collapsed the scale into three groups: "Left" (responses 1–2), "Center" (3–5), and "Right" (6–7). This categorical grouping was used to describe the ideological composition of the sample and to explore heterogeneous treatment effects.

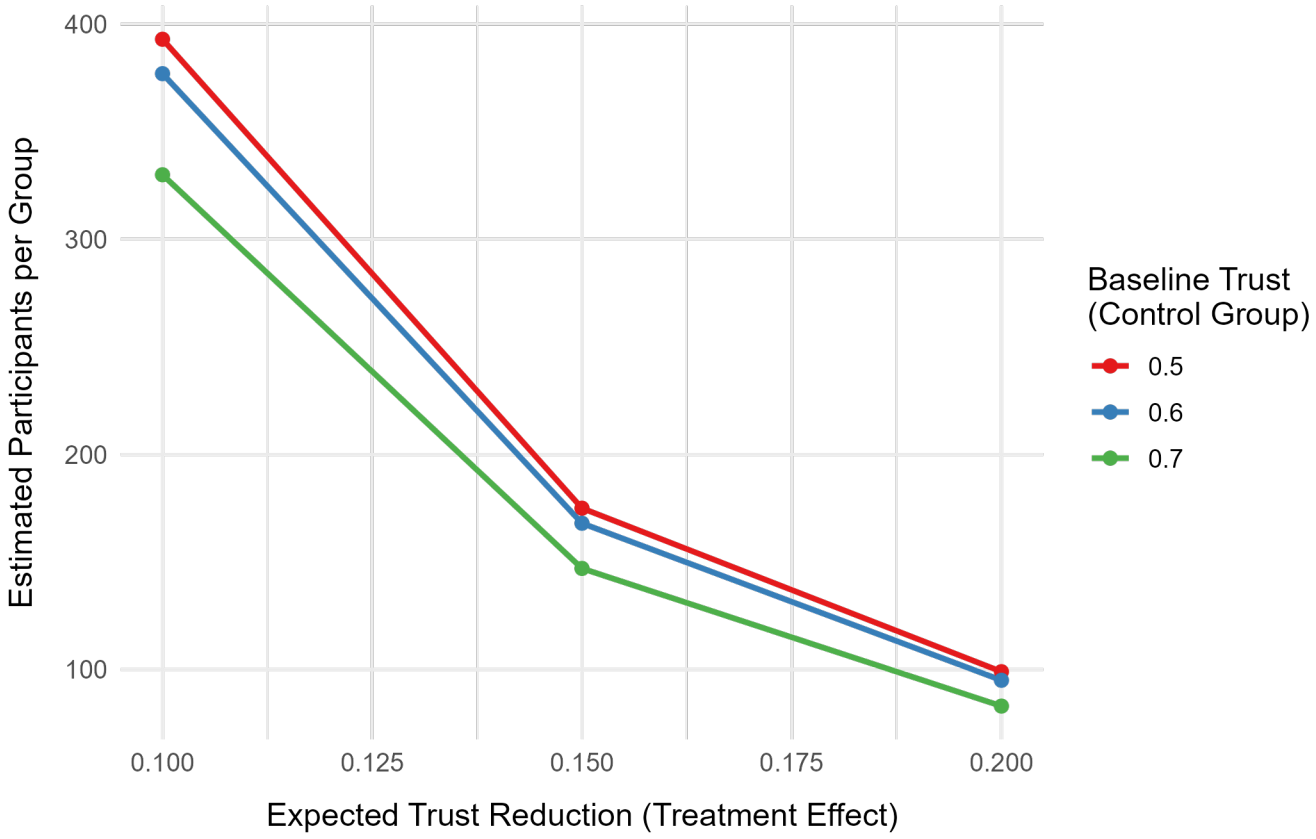

Figure 2: Required Sample Size per Group: This figure shows a naive power analysis. More precisely, it shows the estimated number of participants needed per group (treatment or control) to detect various expected reductions in perceived authenticity of AI-generated content. The x-axis indicates the assumed treatment effect (in percentage points), while each line represents a different baseline level of perceived authenticity in the control group. The plot assumes $\alpha = 0.05$ and a desired power of 0.8.

		Control Group		Process Label Group		Harm Label Group	
Variable		**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Age		31.32	8.70	30.12	8.75	30.60	9.25
Internet Skills		4.32	0.56	4.28	0.60	4.31	0.54
AI Skills		3.45	0.70	3.35	0.77	3.35	0.80
AI Use		0.57	0.39	0.55	0.42	0.52	0.39
Social Media Use		1.10	0.40	1.13	0.40	1.08	0.41
		N	**%**	**N**	**%**	**N**	**%**
Income Group							
	Low	94	32.4	89	30.6	110	37.2
	Middle	143	49.3	116	39.9	129	43.6
	High	36	12.4	61	21.0	43	14.5
	Missing	17	5.9	25	8.6	14	4.7
Political Ideology							
	Left	86	29.7	94	32.3	109	36.8
	Center	192	66.2	190	65.3	179	60.5
	Right	12	4.1	7	2.4	8	2.7
Highest Degree							
	Still in Education	62	21.4	65	22.3	56	18.9
	Secondary	41	14.1	36	12.4	48	16.2
	Tertiary	173	59.7	175	60.1	178	60.1
	No or Other	14	4.8	15	5.2	14	4.7
Gender							
	Female	134	46.2	141	48.5	149	50.3
	Male	152	52.4	145	49.8	140	47.3
	Other	4	1.4	5	1.7	7	2.4
Total N per group		290		291		296	

Table 4: Covariate Balance Across Experimental Groups

ID	Short label	Primary topic	Content type	Germany	International
1	Assange	Politics	Conflict & Geopolitics		✓
2	Cologne protest	Politics	Protest / Collective Action	✓	
3	Georgia protests	Politics	Protest / Collective Action		✓
4	Green Party congress	Politics	Elite Politics	✓	
5	Greta Thunberg	Politics	Conflict & Geopolitics		✓
6	UK police & imams	Politics	Conflict & Geopolitics		✓
7	Neuer red card	Sports	Entertainment / Sports	✓	
8	Assad celebration (Berlin)	Politics	Protest / Collective Action	✓	
9	Putin–Trump dinner	Politics	Elite Politics		✓
10	Söder at McDonald's	Politics	Elite Politics	✓	
11	Trump in safety vest	Politics	Elite Politics		✓
12	Tyson with Palestine flag	Politics	Conflict & Geopolitics		✓

Table 5: Image Content by Topic, Content Type, and Geographic Context

Image	Unlabeled	Harm-Based Label	Process-Based Label
soeder	0.582	NA	NA
georgia	0.556	NA	NA
koeln	0.718	NA	NA
oranienstr	0.744	NA	NA
trump	0.529	NA	NA
neuer	0.629	NA	NA
assange	0.382	0.256	0.254
tyson	0.389	0.248	0.255
greens	0.193	0.110	0.137
imame	0.108	0.070	0.088
greta	0.350	0.244	0.217
putin	0.207	0.129	0.117

Table 6: Average Perceived Authenticity Scores by Image and Labeling Condition

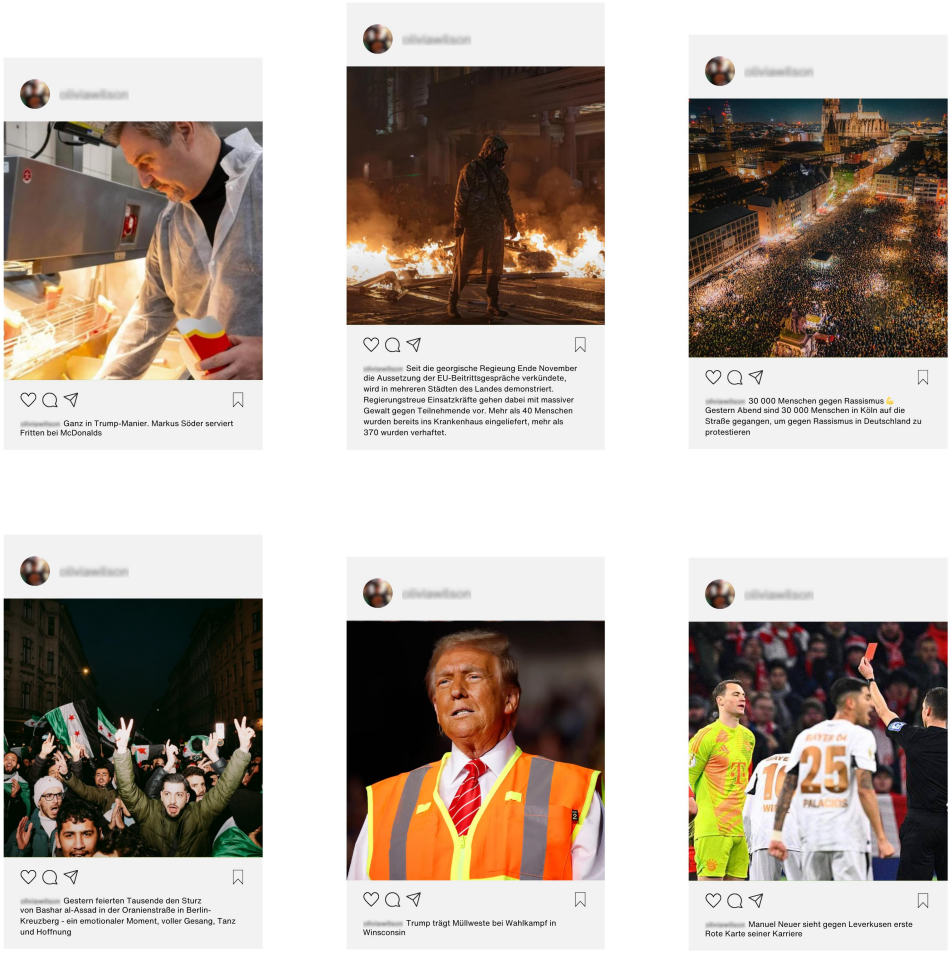

Figure 3: Overview of the six authentic images used in the experiment (Page 1 of 3).

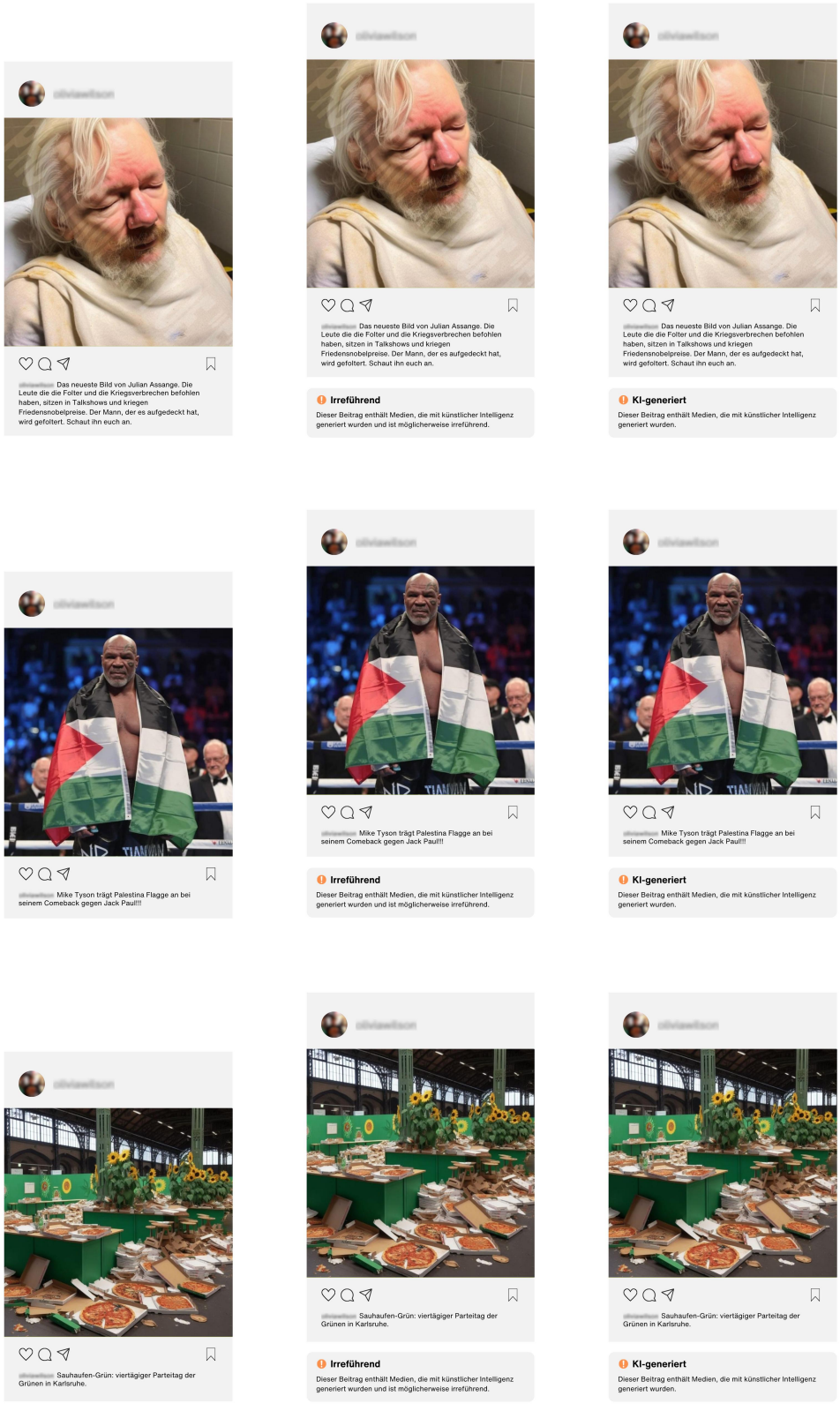

Figure 4: Overview of 3 out of 6 AI-generated or AI-altered image sets used in the experiment (Page 2 of 3).

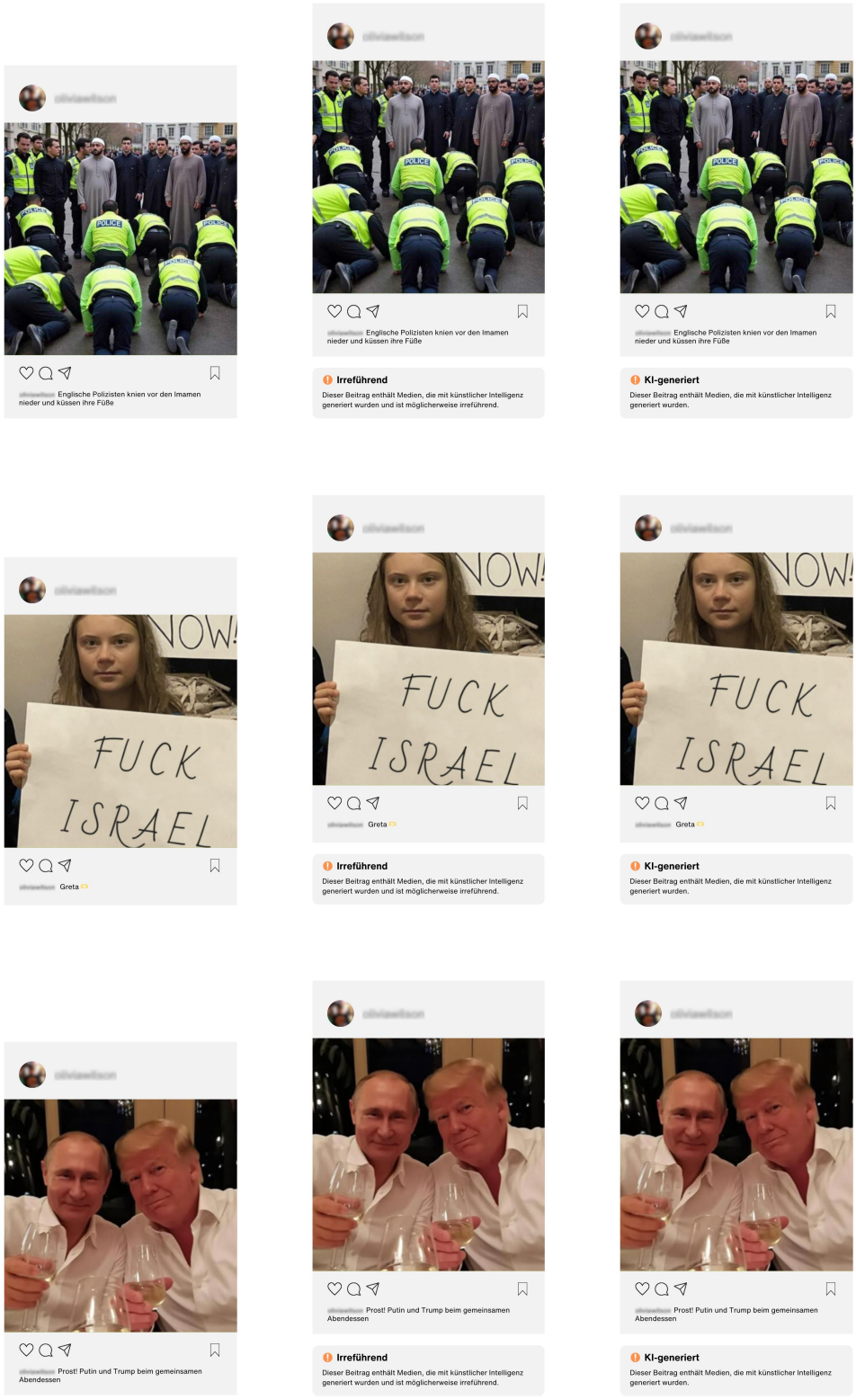

Figure 5: Overview of the remaining 3 out of 6 AI-generated or AI-altered image sets used in the experiment (Page 3 of 3).

	Unadjusted Model		**Preferred Model**	
	Coef.	SE	Coef.	SE
Process Label	-0.094**	0.020	-0.091**	0.019
Harm Label	-0.096**	0.016	-0.093**	0.015
Internet Skills			-0.030*	0.012
AI Skills			-0.004	0.010
AI Use			0.031	0.023
Social Media Use			-0.008	0.017
Age			0.002*	0.001
Low Income			-0.006	0.017
High Income			-0.016	0.015
Missing Income			0.000	0.015
Male			-0.018	0.017
Other Gender			-0.056*	0.021
Secondary Education			0.012	0.022
Still in Education			0.016	0.016
Tertiary Education			0.005	0.017
Left			-0.012	0.011
Right			0.024	0.034
Process × Internet Skills			-0.001	0.018
Process × AI Skills			0.019	0.017
Process × AI Use			-0.009	0.031
Process × Social Media Use			-0.034	0.031
Process × Age			0.000	0.002
Harm × Internet Skills			0.010	0.015
Harm × AI Skills			-0.001	0.021
Harm × AI Use			0.029	0.024
Harm × Social Media Use			-0.012	0.020
Harm × Age			-0.001	0.001
Num. of participants	877		877	
Num. of observations	5,262		5,262	
R^2	0.156		0.168	
Image FEs	Yes		Yes	
Clustered SEs	Yes		Yes	

Table 7: Average Treatment Effects (ATE): full specifications. This table reports the full regression specifications corresponding to the ATE analysis presented in the main text (Table 2), including all pre-specified covariates and interaction terms. Entries report coefficients with clustered standard errors. Standard errors are clustered at the participant and image level. Continuous variables are mean-centered. Statistical significance: * $p < 0.05$; ** $p < 0.01$.

	Model (1) AC Incl.		**Model (2) AC Incl.**	
	Coef.	SE	Coef.	SE
Process Label	-0.093	(0.019)	-0.089	(0.018)
Harm Label	-0.093	(0.016)	-0.090	(0.015)
Internet Skills			-0.033	(0.011)
AI Skills			-0.004	(0.010)
AI Use			0.031	(0.019)
SM Use			-0.007	(0.018)
Age			0.002	(0.001)
Low Income			-0.006	(0.016)
High Income			-0.016	(0.015)
Missing Income			-0.002	(0.015)
Male			-0.016	(0.016)
Other Gender			-0.060	(0.021)
Secondary Education			0.014	(0.019)
Still in Education			0.019	(0.017)
Tertiary Education			0.008	(0.016)
Left (Ideology)			-0.008	(0.011)
Right (Ideology)			0.030	(0.033)
Process × Internet Skills			-0.008	(0.018)
Process × AI Skills			0.023	(0.017)
Process × AI Use			-0.010	(0.027)
Process × Social Media Use			-0.036	(0.035)
Process × Age			0.001	(0.002)
Harm × Internet Skills			0.013	(0.015)
Harm × AI Skills			-0.005	(0.023)
Harm × AI Use			0.044	(0.029)
Harm × Social Media Use			-0.007	(0.019)
Harm × Age			-0.000	(0.001)
Num. Obs.	5,454		5,454	
R^2	0.152		0.167	
Image Fixed Effects	Yes		Yes	
Clustered SEs	Yes		Yes	

Table 8: Robustness Check for ATE: Models including participants who failed attention checks. All models follow the specifications of the preferred models in the main text (see Table 2) but are estimated on the full sample. Standard errors are clustered at the image and participant level. Continuous variables are mean-centered. Statistical significance: * $p < 0.05$, ** $p < 0.01$. Attention Checks included (AC Incl.).

	Unadjusted Model		**Preferred Model**		**AI Interaction Model**	
	Coef.	SE	Coef.	SE	Coef.	SE
Label Exposure	0.018*	0.007	0.019*	0.006	0.023**	0.006
Internet Skills			-0.011	0.012	-0.011	0.012
AI Skills			-0.004	0.008	-0.004	0.008
AI Use			-0.001	0.018	-0.001	0.018
Social Media Use			-0.018	0.011	-0.018	0.011
Age			0.002*	0.001	0.002*	0.001
Low Income			0.009	0.010	0.009	0.010
High Income			0.004	0.009	0.004	0.009
Missing Income			0.002	0.014	0.002	0.014
Male			0.024	0.016	0.024	0.016
Other Gender			-0.056*	0.022	-0.056*	0.022
Secondary Education			0.010	0.025	0.011	0.025
Still in Education			0.007	0.020	0.007	0.020
Tertiary Education			0.005	0.017	0.005	0.017
Left			0.005	0.013	0.005	0.013
Right			0.020	0.024	0.020	0.024
Exposure × Internet Skills			0.006	0.013	0.006	0.013
Exposure × AI Skills			-0.011	0.008	-0.011	0.009
Exposure × AI Use			0.035	0.022	0.035	0.022
Exposure × Social Media Use			0.021	0.019	0.021	0.019
Exposure × Age			-0.002*	0.001	-0.002*	0.001
Exposure × AI Content					-0.012	0.013
Num. of participants	877		877		877	
Num. of observations	8,773		8,773		8,773	
R^2	0.366		0.371		0.372	
Image FEs	Yes		Yes		Yes	
Clustered SEs	Yes		Yes		Yes	

Table 9: Spillover effects on perceived authenticity in unlabeled images (full specifications). This table reports the full set of regression specifications corresponding to the spillover analysis. Columns show the unadjusted model, the preferred model with the full set of pre-specified participant covariates, and an additional model including interaction terms to explore potential heterogeneity. Standard errors are clustered at the participant and image level. Continuous variables are mean-centered. Statistical significance: * $p < 0.05$; ** $p < 0.01$.

	Model (1) AC Incl.		Model (2) AC Incl.		Model (3) AC Incl.	
	Coef.	SE	Coef.	SE	Coef.	SE
Label Exposure	0.0186*	(0.0066)	0.0189*	(0.0062)	0.0254**	(0.0063)
Internet Skills			-0.0141	(0.0124)	-0.0141	(0.0124)
AI Skills			-0.0045	(0.0083)	-0.0045	(0.0083)
AI Use			-0.0054	(0.0188)	-0.0054	(0.0188)
Social Media Use			-0.0185	(0.0117)	-0.0186	(0.0118)
Age			0.0018*	(0.0007)	0.0018*	(0.0007)
Low Income			0.0079	(0.0092)	0.0081	(0.0092)
High Income			0.0062	(0.0092)	0.0063	(0.0092)
Missing Income			0.0023	(0.0141)	0.0024	(0.0142)
Male			0.0267	(0.0164)	0.0267	(0.0165)
Other Gender			-0.0548*	(0.0224)	-0.0549*	(0.0224)
Secondary Education			0.0099	(0.0226)	0.0101	(0.0225)
Still in Education			0.0091	(0.0186)	0.0090	(0.0186)
Tertiary Education			0.0062	(0.0155)	0.0063	(0.0155)
Left (Ideology)			0.0070	(0.0132)	0.0070	(0.0132)
Right (Ideology)			0.0240	(0.0240)	0.0239	(0.0239)
Exposure × Internet Skills			0.0127	(0.0124)	0.0127	(0.0124)
Exposure × AI Skills			-0.0099	(0.0086)	-0.0098	(0.0086)
Exposure × AI Use			0.0375	(0.0229)	0.0374	(0.0229)
Exposure × Social Media Use			0.0209	(0.0180)	0.0209	(0.0180)
Exposure × Age			-0.0015*	(0.0007)	-0.0015*	(0.0007)
Exposure × AI Content					-0.0168	(0.0124)
Observations	9,100		9,100		9,100	
R^2	0.361		0.366		0.367	
Image Fixed Effects	Yes		Yes		Yes	
Clustered SEs	Yes		Yes		Yes	

Table 10: Robustness Check for Spillover Model (Effect of Label Exposure): Models including participants who failed attention checks. All models follow the specifications of the preferred models in the main text (see Table 3) but are estimated on the full sample. Standard errors are clustered at the image and participant level. Continuous variables are mean-centered. Statistical significance: * $p < 0.05$, ** $p < 0.01$. Standard errors clustered at participant and image level. Attention Checks included (AC Incl.).

Variable Name	Type	Description
trust_score	numeric (0–1)	Participant's trust rating for an image (0 = no trust, 1 = complete trust). This refers to perceived authenticity.
treatment	categorical	Treatment assignment: "baseline", "ai_label", or "misleading_label."
is_labeled	dummy (0/1)	Shown image had a label (1 = yes, 0 = no).
process_label_actual	dummy (0/1)	Whether the image had a process-based label ("AI-generated").
harm_label_actual	dummy (0/1)	Whether the image had a harm-based label ("Misleading").
process_group	dummy (0/1)	Participant assigned to process-based label group.
harm_group	dummy (0/1)	Participant assigned to harm-based label group.
is_ai_generated	dummy (0/1)	Whether the shown image was AI-generated or AI-altered.
exposed_to_label	dummy (0/1)	Whether the participant had seen at least one labeled image before the current one.

Table 11: Codebook: Outcome and Treatment Variables

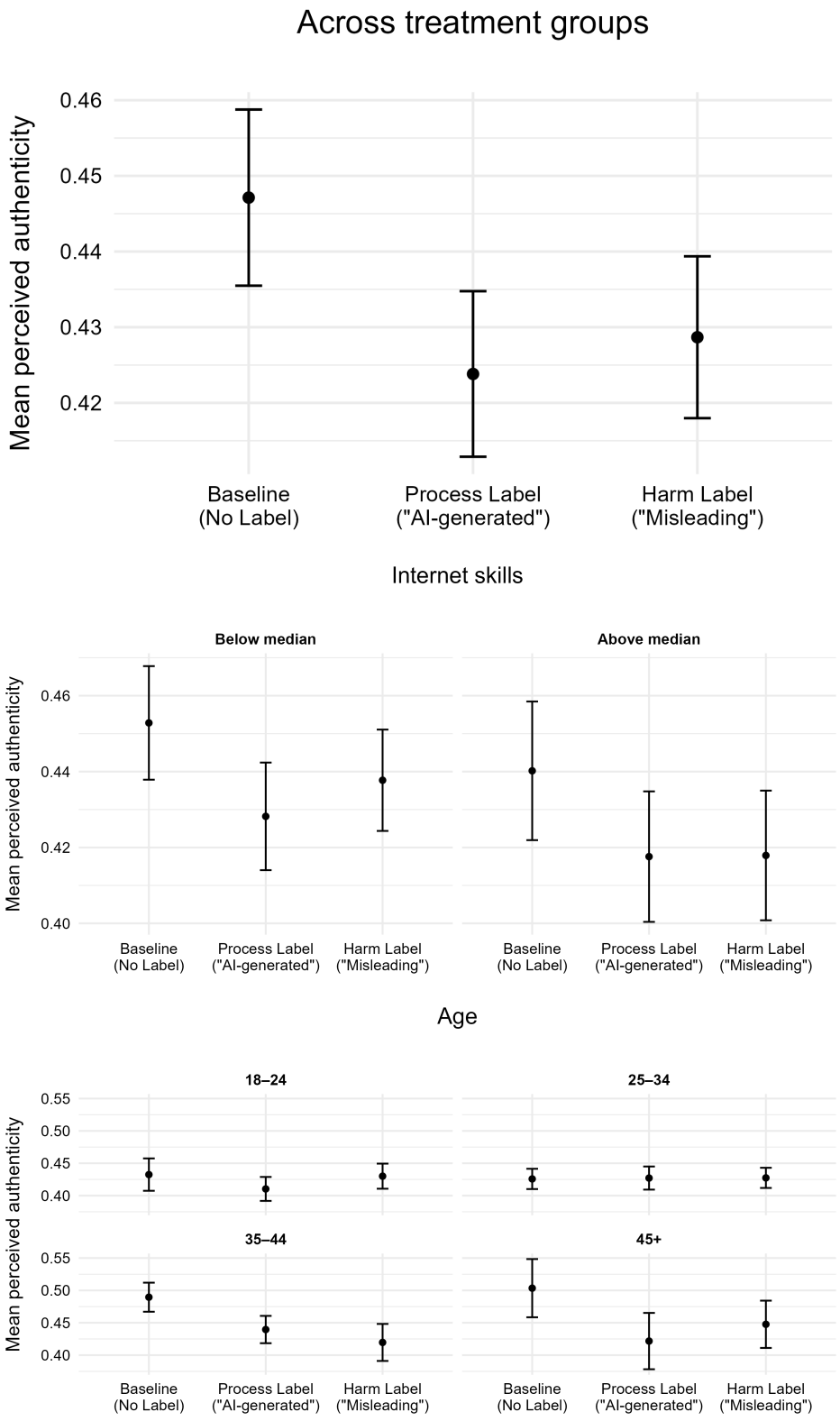

Figure 6: Perceived authenticity across conditions and participant subgroups. Panel 1 shows participant-level mean perceived authenticity across experimental conditions with 95% confidence intervals. Panel 2 shows the same estimates stratified by internet skills (below vs. above the median). Panel 3 shows the same estimates stratified by age group.

Variable Name	Type	Description
gender	categorical	Participant's gender: Male, Female, Other.
age_c	numeric	Participant's age, mean-centered.
education_level	categorical	Highest school qualification: Low, Medium, High, Other (German system).
highest_degree	categorical	Collapsed highest degree: Tertiary, Secondary, Still in Education, No/Other Degree.
income_relative	categorical	Relative household income group: Low, Middle, High, Missing.
political_ideology	categorical	Political ideology group: Left (1–2), Center (3–5), Right (6–7). (Kozyreva et al. 2021; Yan, Schroeder, and Stier 2022)

Table 12: Codebook: Demographic and Socioeconomic Covariates

Variable Name	Type	Description
internet_skills_c	numeric	Global Internet Skills Index (1–5), mean-centered. (Deursen 2010)
ai_skills_c	numeric	AI Skills Index (1–5), mean-centered. (Lintner 2024; Lee and Park 2024)
ai_use_c	numeric	Frequency of generative AI platform use (0–4 scale), mean-centered.
social_media_use_c	numeric	Frequency of social media platform use (0–4 scale), mean-centered.

Table 13: Codebook: Digital Skills and Platform Usage Covariates

Variable	Mean	SD	Min	Max
Age	30.68	8.91	18	72
Household Income (Monthly, €)	3,124	2,598	250	30,000
Internet Skills (1–5)	4.30	0.57	1.56	5
AI Skills (1–5)	3.39	0.74	1	5
AI Use (0–4 scale)	0.55	0.40	0	2.9
Social Media Use (0–4 scale)	1.10	0.40	0.07	3
Political Ideology (1–7)	3.16	1.25	1	7

Table 14: Descriptive Statistics for Key Variables

Moderator	Subgroup	Label	Diff. from ATE	SE	p (raw)	p (BH-FDR)
AI skills	NA	Harm label	0.018	0.019	0.349	0.968
AI skills	NA	Process label	-0.007	0.025	0.764	0.965
AI use	NA	Harm label	0.007	0.019	0.720	0.968
AI use	NA	Process label	-0.005	0.029	0.865	0.965
Age	NA	Harm label	-0.002	0.023	0.920	0.968
Age	NA	Process label	-0.014	0.028	0.611	0.965
Education	Other	Harm label	-0.069	0.031	0.024	0.586
Education	Secondary	Process label	-0.052	0.033	0.114	0.965
Gender	Other	Harm label	-0.089	0.049	0.068	0.663
Gender	Other	Process label	0.100	0.074	0.174	0.965
Ideology	Right	Harm label	-0.015	0.057	0.787	0.968
Ideology	Right	Process label	0.020	0.037	0.582	0.965
Income	Missing	Harm label	-0.023	0.061	0.712	0.968
Income	Low	Process label	-0.011	0.030	0.711	0.965
Internet skills	NA	Harm label	0.014	0.011	0.180	0.968
Internet skills	NA	Process label	-0.005	0.022	0.807	0.965
SM use	NA	Harm label	0.005	0.017	0.785	0.968
SM use	NA	Process label	-0.023	0.031	0.455	0.965

Table 15: Subgroup Differences from the Average Treatment Effect with False-Discovery Correction. For each moderator and label type, the table reports the subgroup exhibiting the largest absolute deviation from the overall average treatment effect.

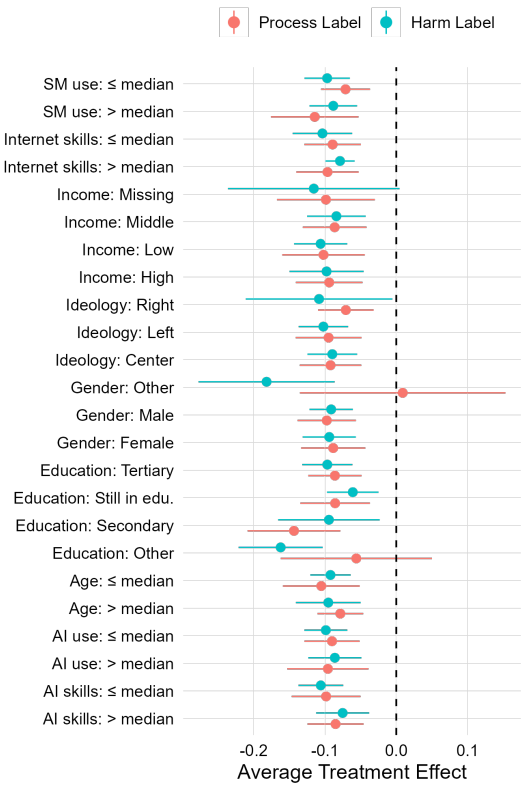

Figure 7: Subgroup average treatment effects (ATE) of the Process and Harm labels on the perceived authenticity scale (ranging from 0-1) across categorical and median-split continuous moderators. Points are coefficient estimates; whiskers are 95% CIs with SEs clustered by participant and image. Estimates are from separate regressions within each subgroup with image fixed effects and the pre-specified controls, excluding the subgroup variable. Sample restricted to AI-generated images.

Figure 8: Subgroup spillover effects of prior label exposure on *perceived authenticity scale* (ranging from 0-1) for *unlabeled* images. Points are coefficient estimates; whiskers are 95% CIs with SEs clustered by participant and image. Continuous moderators are median-split; models are re-estimated within each subgroup with image fixed effects and the pre-specified controls, excluding the subgroup variable.