# C2PA Provenance Labels Increase Trust in News Platforms Across Western Countries

**Christoph Trattner**[1], **Svenja Lys Forstner**[1], **Alain D. Starke**[1,2], **Erik Knudsen**[1]

[1]MediaFutures: Research Centre for Responsible Media Technology & Innovation, University of Bergen, Lars Hilles gate 30, 5008 Bergen, Norway

[2]Amsterdam School of Communication Research, University of Amsterdam, 1012 WP Amsterdam, Netherlands

christoph.trattner@uib.no, svenja.forstner@uib.no, alain.starke@uib.no, erik.knudsen@uib.no

## Abstract

Misinformation and disinformation threaten global public trust in news media. Generative AI exacerbates mistrust by making it difficult to distinguish authentic images from AI-generated ones. This study examines whether accompanying images with C2PA (Coalition for Content Provenance and Authenticity) provenance labels can restore trust. C2PA is an open standard that cryptographically secures and describes a media file's origin and editing history. We conducted an online experiment with $N = 6,114$ participants, reflecting audiences of six major news sources in the US, UK, and Norway. Each participant evaluated six news article previews with images, either accompanied by a provenance label (three levels of detail) or not. Presenting provenance metadata to participants significantly improved their perceptions of an image's transparency and credibility, and also increased feelings of trust in a presented news source. These results show that verifiable provenance makes visual content more inspectable and strengthens brand trust. By adopting C2PA or similar frameworks, news organizations can counter AI-generated disinformation and improve audience trust.

## Introduction & Related Work

Determining the authenticity of digital images has become increasingly challenging. Recent advances in generative AI have made it easier and more affordable than ever to create, adapt, and manipulate visual media to look remarkably authentic (Langguth et al. 2021; Hwang and Oh 2023). Platforms such as ChatGPT demonstrate the ability of generative AI to produce convincing synthetic photos and short videos (Bansal et al. 2024). Consequently, distinguishing authentic from manipulated content is difficult for users of digital news platforms (Newman et al. 2024). This is illustrated by research findings that individuals are generally less accurate in identifying authentic and true information than in identifying false news (Pfänder and Altay 2025).

Generative AI has intensified global concerns about misinformation and disinformation (Cavaciuti-Wishart et al. 2024; Ipsos 2023). According to a survey by Ipsos and UNESCO, 94% of 8,000 international respondents reported experiences of being misled by online disinformation (Ipsos 2023). Such experiences with 'fake content' align with declining trust in media across many nations (Egelhofer and Lecheler 2019; Strömbäck et al. 2020). For instance, Gallup found that only 32% of U.S. respondents trusted mass media in 2023, marking an 8% drop since 2020 (Brenan 2023). Similarly, the Reuters Institute Digital News Report shows decreasing media trust in most of the 47 countries studied, including Greece and Hungary with trust levels of just 23%. Only a few countries, such as Norway (55%) and Finland (69%), have maintained relatively stable trust scores (Newman et al. 2024). The increasing proliferation of AI-generated media further threatens trust in news media, undermining the public's ability to establish shared social beliefs and concerns (Strömbäck et al. 2020; Hanitzsch, Dalen, and Steindl 2018; Coleman 2012).

Given these challenges, news users could greatly benefit from tools that help assess the credibility of content. One promising approach is the implementation of provenance labels, which clearly indicate to what extent images have been edited or generated by AI (Nsoesie and Ghassemi 2024; Wittenberg et al. 2024). An emerging framework designed to improve image transparency is the Coalition for Content Provenance and Authenticity (C2PA), closely associated with initiatives such as Project Origin[1] and the Content Authenticity Initiative[2]. C2PA enables one to trace the history of origin and modification of digital media through verifiable metadata, including the history of editing, the authorship, and spatio-temporal details (Coalition for Content Provenance and Authenticity 2025). This approach is analogous to the transparency methods used in AI systems, where clarifying the mechanisms of algorithms, which are often perceived as opaque "black boxes", has been shown to foster greater user trust (Bansal et al. 2024; Liu 2021; Zerilli, Bhatt, and Weller 2022).

Integrating C2PA provenance information into news articles can offer concrete methods to improve credibility, transparency, and accountability (Strömbäck et al. 2020; Fisher 2016). This aligns with broader recommendations to make verification processes integral to journalism (Kovach and Rosenstiel 2021; Silverman 2020). Recent efforts to adopt provenance labels based on C2PA are emerging, although empirical peer-reviewed studies that evaluate their effectiveness remain scarce. Feng et al. (2023) explored the use of



[1]https://www.originproject.info/ (accessed April 25, 2025).

[2]https://contentauthenticity.org/ (accessed April 25, 2025).

C2PA labels in a social media context and found that provenance labeling helped users detect manipulative content. However, it also reduced trust in credible sources, suggesting challenges in differentiating between source and provenance credibility (Tsfati and Cappella 2003). Similarly, existing literature has primarily examined textual rather than image-based provenance labels. For example, Martel and Rand (2024) demonstrate the positive influence of textual fact-checking labels on trust toward fact checkers.

BBC Research & Development has recently piloted the implementation of "Content Credentials". These are image-based provenance labels, which have been used in a small number of public articles (Astier and Avagnina 2024), as well as in internal studies (Monday and Strappelli 2024; Marcus 2024). One study shows that the use of Content Credentials labels on the BBC website increases trust among 83% of participants, while 96% consider the Content Credentials to be useful (Monday and Strappelli 2024). Although they suggest with their findings that trust is fostered by increasing transparency and accountability through the presented provenance information (see also Strömbäck et al. (2020) and Fisher (2016)), this is not validated by measuring such concepts. Moreover, the described studies are not peer reviewed and apply only to a single platform, focusing on trust in the BBC News brand (Monday and Strappelli 2024).

We address these gaps by offering a robust investigation of the impact of C2PA provenance labels on source trust. On the one hand, we verify the preliminary research by inquiring perceived transparency of provenance labels, which is shown by Feng et al. (2023) in their studies on provenance labels on social media. On the other hand, we contextualize source trust within a broader theoretical framework of news media trust by Strömbäck et al. (2020), focusing on the perceived credibility of images and its underlying dimensions of fairness, accuracy, and bias (Gaziano and McGrath 1986; Meyer 1988). Strömbäck et al. (2020) consider the role of news media as an entity that informs the public, which supports democratic processes as a result (Helberger 2019; Holbert 2005). In doing so, one can differentiate between aspects of news at different levels of aggregation, ranging from news media in general to specific media content (Strömbäck et al. 2020). C2PA technology, operationalized as provenance labels, could affect how individual citizens consume media content and news stories. One of the transactions in news consumption is accepting the veracity of the information presented (Strömbäck et al. 2020), deeming the accuracy and unbiasedness of the information to be important determinants of news credibility and consumption (Coleman 1998; Mayer, Davis, and Schoorman 1995).

To conceptualize perceived image credibility, we follow research on believability and credibility indices that show how trust is merely one of credibility's subdimensions (Gaziano and McGrath 1986; Meyer 1988; West 1994). Gaziano and McGrath (1986) have identified sixteen subdimensions of perceived credibility of news media, which were later reduced to five items that can be applied to the different levels of news trust analysis (Meyer 1988; Strömbäck et al. 2020; West 1994). Some of these items can be attributed to text-based content, such as factual trust and completeness, but also dimensions that can relate to image credibility, such as fairness, unbiasedness, and information accuracy. In this paper, we will focus on *image credibility* as a phenomenon of media content, and use it to predict trust in a news source or brand, which is set at a higher level of aggregation than media content (Strömbäck et al. 2020). We note that empirical examinations of the relation between media content and trust in media brands are rare (Chan-Olmsted and Kim 2023; Strömbäck et al. 2020).

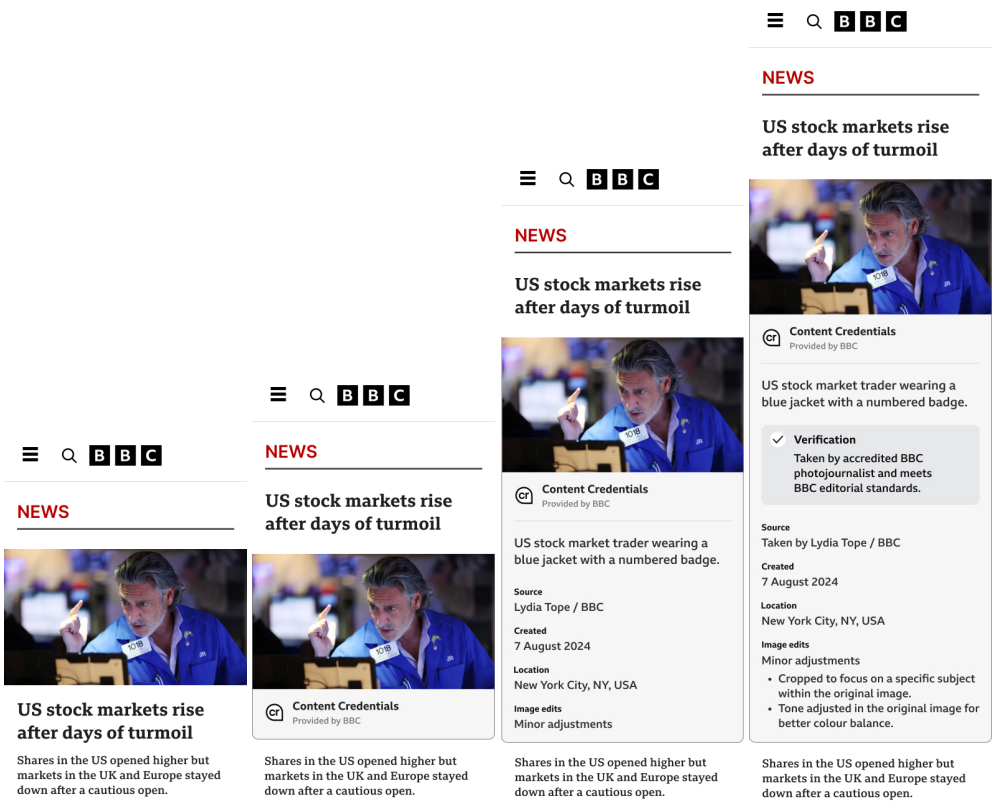
Furthermore, we go beyond the evaluation of a single media brand. We investigate the effects of provenance labels for six international news sources from the UK, US, and Norway, representing varying baseline levels of trust (high and low). For instance, we compare effects on highly trusted sources such as BBC News and less trusted counterparts such as The Sun (Newman et al. 2024). As we investigate the effects on different levels of news media, including individual brands (i.e., source trust) and media content (i.e., image transparency and image credibility), we hypothesize that transparency and perceived image credibility serve as pathways for increasing trust, where we will specifically examine mediating effects between image provenance, perceived credibility (i.e., both level of media content) and source trust (i.e., level of media brands). Furthermore, we analyze whether trust responses vary by country, and whether the effectiveness of provenance transparency depends on baseline trust levels in respective news markets.

Our study operationalizes C2PA provenance information across three increasingly detailed levels, ranging from minimal verification signals to comprehensive metadata disclosure (see Figure 1). While greater detail is generally expected to enhance transparency and trust, preliminary BBC research indicates a potential "overshoot", suggesting simpler labels may sometimes be more effective (Monday and Strappelli 2024; Halford 2024; Marcus 2024). Our research thus seeks to test this assertion through the following research question:

> *How does displaying provenance information for news article images affect source trust, image credibility and transparency in three countries (Norway, UK, US), and does this effect vary between low-trust and high-trust sources?*

Given varying levels of media trust across the UK, US, and Norway (Newman et al. 2024), the study examines whether similar patterns can be observed across the selected countries and news sources. We also report whether the granularity of provenance details directly corresponds to increases in transparency, image credibility, and trust, particularly across intermediate and high levels of provenance information. In addition, our goal is to investigate whether the baseline trust level of a news outlet influences the effectiveness of provenance information in enhancing trust. Specifically, we want to determine whether less trusted outlets experience a larger benefit from displaying provenance details.

In addition to baseline trust, we also examine the moderating role of user familiarity. Prior research has found that familiar sources are generally met with higher trust and are preferred over unfamiliar ones (Peterson and Allamong




(a) Lvl 0 [No CC] (b) Lvl 1 [Logo Only] (c) Lvl 2 [Source CC] (d) Lvl 3 [Full CC]

Figure 1: Graded presentation of image provenance metadata in news-article mockups. Four interface variants display incremental levels of provenance information alongside a BBC mobile news article: *(a) Level 0 [No CC]*, baseline with no provenance metadata; *(b) Level 1 [Logo Only]*, adds a compact "Content Credentials" (CC) badge indicating provenance availability; *(c) Level 2 [Source CC]*, expands the badge to show key metadata (photographer, creation date, location, edit note); and *(d) Level 3 [Full CC]*, full provenance panel including verification statement, structured source attribution, detailed metadata and explicit edit history. These variants were evaluated to determine the impact of provenance disclosure on reader trust and credibility judgments of journalistic images.

2022; Pennycook and Rand 2019), making source familiarity a relevant factor when assessing the impact of provenance labels on trust. As we consider C2PA provenance labels to support trust at the media content level (Strömbäck et al. 2020), we hypothesize that the benefits of provenance labels are greater for users who are unfamiliar with the source, and smaller for regular users. This is supported by preliminary findings from the BBC, which observed increased trust among non-users when BBC news content included provenance labels (Monday and Strappelli 2024). We aim to investigate whether a similar effect occurs across a variety of news sources.

## Methodology

### Overview

Our research question was examined through an online experiment involving participants (N = 6,114) from Norway, the UK, and the USA. Each participant viewed six news articles, including a headline, introductory text, and an accompanying image with or without an image provenance label (see Figures 1a-1d). All participants evaluated articles from their own country. Half of the articles presented were sourced from low-trust media sources (The Sun [UK], ABC Nyheter [NOR], Huffington Post [USA]) and half from high-trust sources (BBC News [UK], NRK [NOR], CBS News [USA]). Participants rated trust in each news source, evaluated the perceived credibility of images, and the perceived transparency of the provided information. Materials can be found through our repository: https://anonymous.4open.science/r/C2PA-3-country-study/.

### Research Design

We conducted an online experiment designed to allow comparison with previous preliminary research (Marcus 2024; Monday and Strappelli 2024). The study employed a 4 (provenance levels) × 2 (low-trust vs. high-trust source) mixed research design. The provenance labels were integrated into the previews of the news articles, comprising a

headline, an image, and an introductory paragraph. Figures 1a through 1d illustrate the layouts of specific articles and the levels of provenance. The degree of provenance detail from the C2PA content credentials (CC) presented beneath each image was varied across four conditions: no credentials (baseline), CC logo only (Level 1), CC with source information including image description, source, date of creation, and location (Level 2), and full CC details including verification practices and a specific description of image edits (Level 3); all based on a pilot study (Marcus 2024).

Participants were randomly assigned to two distinct provenance label conditions, each associated with either a low-trust or high-trust news source, predetermined based on the participant's country of residence (e.g., BBC News and The Sun for participants residing in the UK). Although some participants were inadvertently assigned twice to the same provenance condition, we retained them in our sample. Our random-effects regression analysis could account for such nested participant data.

Each participant encountered two sets of stimuli, each comprising three articles. One set featured content from a low-trust source, the other content from a high-trust source. These sources were selected based on their trust ratings from the most recent Reuters Institute Digital News Report at the time (Newman et al. 2024). The study included participants from Norway, the UK, and the USA. For each country, we selected sources representing the highest and lowest levels of trust, except in Norway. There, we excluded the lowest-rated source, "Document.no", due to its documented history of spreading misinformation, conspiracy theories, and islamophobic content (Faktisk.no 2021; Gardell 2014), opting instead for the source with the next lowest trust rating.

The selected sources were NRK (80% trust, 10% distrust) and ABC Nyheter (44% trust, 22% distrust) in Norway; CBS News (52% trust, 28% distrust) and HuffPost (39% trust, 32% distrust) in the USA; and BBC News (62% trust, 22% distrust) and The Sun (15% trust, 63% distrust) in the UK.

## Procedure

The experimental procedure was designed in line with the preliminary study conducted by the BBC (Monday and Strappelli 2024). All participants provided their informed consent before the study. Each participant followed an identical survey structure, differing only in the stimuli presented. The survey was designed according to the ethical standards of the Research Council of Norway and did not require further review. It consisted of four primary sections.

**Part 1** captured general information about participants' online news consumption habits, trust in their national media landscape (i.e., general media trust), and attitudes toward the two selected sources. In addition, this section queried participants' interest and self-assessed knowledge about financial and economic topics.

**Part 2** evaluated the perceived clarity and comprehensibility of the provenance information in the presented stimuli. Taken together, these items represented perceived image transparency. Participants completed these assessments for each of the three initial stimuli.

**Part 3** focused explicitly on source trust. In addition, it inquired on the perceived accuracy, fairness and perceived bias of the images presented by that source (i.e., items for perceived credibility). Participants also reported factors that influenced their evaluations. Note that parts 2 and 3 were repeated for a second set of three stimuli.

**Part 4** provided a debriefing, explaining the study's objectives and revealing differences in stimuli presentation across experimental groups. The survey concluded with an open-ended prompt inviting feedback on the provenance information's presentation.

## Materials

The survey was adapted to the specific contexts of each participating country. Modifications included translating the original English questionnaire into Norwegian and customizing questionnaire items to reflect each country's unique news media landscape. Regarding language adjustments, only minor wording variations were introduced in the English questionnaires to account for differences between US and UK English. The Norwegian translation was reviewed and verified by two native Norwegian speakers. Additionally, the selection options were adjusted from the original UK version by utilizing data from YouGov[3] for the US context and Medietall[4] for Norway.

The survey stimuli consisted of typical news headlines presented alongside an introductory sentence and a cover image, all formatted for mobile viewing. Participants were exposed to varying levels of image provenance information, divided into four distinct groups (see Figure 1). Level 0 did not include provenance details. Level 1 provided basic provenance information, such as the logo and the name of the information provider. Level 2 expanded upon this by incorporating source, creation date, location details, and information on image edits. Finally, Level 3 included comprehensive verification details, an image description, and a detailed account of previous edits.

The provenance information was sourced primarily from respective news sources and original image providers. In cases where some details were not publicly accessible, supplemental information was generated to ensure consistency and completeness across stimuli. Due to limited verifiability regarding image edits, uniform placeholder content was employed in edit sections. Image descriptions for Levels 2 and 3 were sourced directly or slightly modified from available descriptions; otherwise, descriptions were generated when unavailable.

To minimize the risk of polarization, the stimuli featured neutral topics such as finance, economics, and stock market developments. However, each news source's set included one slightly polarizing article to allow comparative analyses. Articles selection was also based on the availability of image metadata, reducing the need for manual supplementation. Only stock or editorial images were included; gen-

[3]https://today.yougov.com/ratings/entertainment/popularity/news-websites/all (accessed April 25, 2025).

[4]https://www.medietall.no/index.php?liste=persontall (accessed April 25, 2025).

erated images were deliberately excluded to ensure consistency in provenance information. All articles were recent (at the time the study was conducted), published from April 2024 onwards, and formatted for mobile devices, aligning with predominant news consumption habits (Dunaway, Johanna and Searles, Kathleen 2023).

## Participants

A representative sample of N = 6,114 adult participants completed our study. To allow for cross-country comparisons and diversity in media landscapes, we recruited residents from the USA (n = 2,053), UK (n = 2,044), and Norway (n = 2,017). This reflected varying levels of trust. For example, citizens in the US tend to have a lower and decreasing level of trust in the news media compared to Norway (Hanitzsch, Dalen, and Steindl 2018; Newman et al. 2024).

Participants were recruited using the YouGov market research and data analysis platform[5], which had a large base of respondents in all selected countries. We chose to collect a large sample per country ($n \simeq 2,000$) to ensure representativeness. The sample size particularly applied to the large population of the USA, but also to be able to detect small effect sizes in a 4x2-mixed design. All participants were randomly assigned to two conditions of image provenance (No CC: $n = 3,047$; Logo Only: $n = 3,044$; Source CC: $n = 3,073$; Full CC: $n = 3,064$), and both conditions of outlet trust (low vs high).

Data collection was carried out in two rounds for each country. This concerned a period of 11 days (28 October - 7 November 2024) in the US, 12 days (28 October - 8 November 2024) in the UK, and 26 days (30 October - 25 November 2024) in Norway. The respondents received points for their participation, corresponding to 750 points in the US (where 25,000 points can be exchanged for a $15 gift card), 50 points in the UK (with 5,000 points to be exchanged for a £50 gift card), and 75 points in Norway (where 1,500 points can be exchanged for a NOK100 gift card). All participants complied to a default informed consent form set up by YouGov. No personal data was stored during data collection, while open-ended answer fields were anonymized if needed.

The demographic data collected included age ($M = 48.66$, $SD = 16.77$), gender (50.79% female), region, education, household income, urbanization, occupation, and political views. Demographics were mostly similar for individual countries, such as for age: Norway ($M = 48.57$, $SD = 17.19$), the UK ($M = 48.63$, $SD = 18.11$), and the USA ($M = 48.77$, $SD = 17.68$). The sample was well balanced regarding age and gender in all countries. Similarly, in terms of education, between 33.9% - 41.3% of the participants in each country held at least a Bachelor degree. Regarding political viewpoints, in the UK and US, a majority of 30.7% positioned themselves in the middle between left and right. 30.2% were leaning towards the right side and 31.3% to the left. In Norway, 18.8% considered themselves as centrist. A majority of 37.6% was either moderate right-wing or right-wing, while 28.1% defined themselves as moderate left-wing or left-wing. Note that although we shared the collected dataset, the data cannot be used to identify individual participants. Hence, we did not collect 'personal data'.

[5]https://yougov.co.uk (accessed April 25, 2025).

## Measures

The independent variable of image provenance was operationalized in two ways. On the one hand, we used it as a continuous variable with values 0-3 for the four levels of provenance, assuming linear effects. On the other hand, we operationalized it as a set of dummy variables, in which the three label conditions (level 1-3) were compared to the no-CC baseline (level 0).

The main dependent variable was source trust, focusing on individual media brands. It was derived from the media trust framework of Strömbäck et al. (2020) that focused on five layers of media trust. These related direct measurements of trust to "information coming from news media", focusing here on trust in a media brand: "I generally trust information from [news source]", measured on a 7-point Likert scale. This was used to measure trust evaluations at the start of the questionnaire, as well as following each set of stimuli for specific sources. In addition, we adapted one item from Strömbäck et al. (2020) to inquire about a participant's general media trust: "I generally trust information from the news media in the UK", measured on a 7-point Likert scale. This was used to explore moderating effects of image provenance on source trust.

For measuring the perceived credibility of presented images, we used the proposed items of measuring media trust by Strömbäck et al. (2020), based on the Meyer (1988) credibility index and credibility items (Gaziano and McGrath 1986). As questionnaire items related to factuality and completeness arguably did not apply to images, we only included questionnaire items that inquired on the fairness, accuracy and unbiasedness of media content, which were all important determinants of trust (Strömbäck et al. 2020). We designed 3 questionnaire items based on Gaziano and McGrath (1986) to assess how fair, accurate, and biased our images were in the context of economics and finance news. These items were found to form a single construct, which was labelled as 'perceived image credibility' ($\alpha = 0.858$).

Items to measure image transparency were based on questionnaire items from Feng et al. (2023). Using a set of 7 items (measured on 7-point Likert scales), we inquired on the perceived clarity regarding the image's content (i.e., "It was clear what the image was showing"), creator, source, creation date, verification, potential edits to the image, and where to find further information about the image. These items reliably formed a single construct ($\alpha = 0.896$).

We also used demographic factors to examine possible interaction effects with image provenance, obtained from YouGov. This included age (continuous), gender (dichotomous: male or female), and political orientation (participants were asked to place themselves on a 10-point scale, ranging from 'left' to 'right').

# Results

## Source Trust

We first examined how provenance information affected trust towards an individual news source or brand. We employed four random-effects linear regression models clustered at the user level, analyzing a participant's source trust evaluation following each source. We found that increasing levels of provenance information significantly increased source trust, both as an ordinal measure ($\beta = 0.17$, $p < 0.001$) and when compared individually against the no-CC baseline (Level 1: $\beta = 0.13$, $p < 0.01$; Level 2: $\beta = 0.29$, $p < 0.001$; Level 3: $\beta = 0.50$, $p < 0.001$; see Model 1.1 in Table 1). We ran a similar analysis using a dependent variable that measured the pre-post differences in trust and obtained similar results (Level 1: $\beta = 0.096$; Level 2: $\beta = 0.31$; Level 3: $\beta = 0.48$; all $p < 0.001$; not included in Table 1 for brevity). Together, these findings highlight that more detailed provenance labeling significantly increased the perceived trustworthiness of news sources.

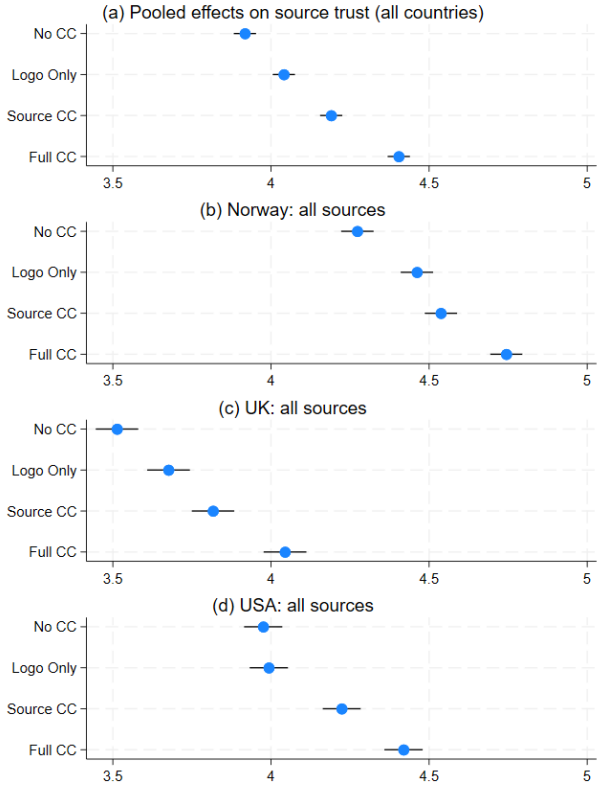



Figure 2: Impact of provenance information on perceived source trustworthiness (pooled effects). Mean trust ratings (dots) and 95% confidence intervals (bars) for four provenance-cue conditions: No CC, Logo Only, Source CC and Full CC. (a) Pooled data across all countries; (b–d) country-specific effects for Norway, the UK and the USA.

We visualize the overall effects in Figure 2, along with country-specific effects. While the overall effect per provenance level shows a significant increase in source trust with each step (see Figure 2a), there were some country-specific differences. Most notably, only showing a CC logo did not lead to an increase in trust for US-sourced media (see Figure 2d). However, the full CC (level 3 of provenance) attained the highest level of source trust in all countries.

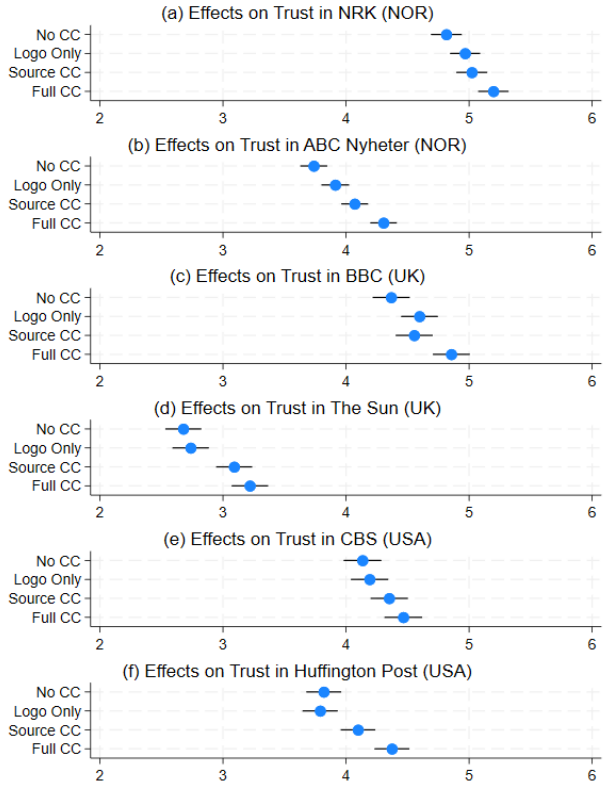



Figure 3: Effects on source trust in news sources across three countries, Norway (NOR) [a,b], United Kingdom (UK) [c,d] and United States of America (USA) [e,f]. Mean trust ratings (dots) and 95% confidence intervals (horizontal lines) are shown for the four levels in our labeling condition variable: No CC, Logo Only, Source CC, Full CC; across each source: NRK and ABC Nyheter (Norway), BBC and The Sun (United Kingdom), and CBS and Huffington Post (United States). Trust was measured with the item 'I generally trust information from [source-name]', on a 7-point Likert scale.

**High-trust vs. Low-trust Sources** We further examined how provenance details affect trust in low versus high-trust media (Model 1.2, Table 1). We again found that provenance information positively influenced trust overall ($\beta = 0.17$, $p < 0.001$), but also observed that low-trust sources were rated significantly lower ($\beta = -1.05$, $p < 0.001$), which was in line with the findings of the Reuters Digital News Report (Newman et al. 2024). Importantly, the interaction showed that provenance information yielded stronger trust gains for low-trust sources ($\beta = 0.053$, $p = 0.037$; Figure 3). This suggested that if a source suffered from a relatively low level of public trust, the addition of C2PA prove-

Table 1: Three multilevel linear regression models predicting *source trust*, based on the presented provenance level and whether a source had a low baseline level of trust. The results of Models 3.2-3.4 in Table 3 suggests full mediation between provenance and source trust through image credibility (Baron and Kenny 1986). $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

| | Model 1.1<br>$\beta$ (S.E.) | Model 1.2<br>$\beta$ (S.E.) | Model 1.3<br>$\beta$ (S.E.) | Model 1.4<br>$\beta$ (S.E.) |
|---|---|---|---|---|
| Provenance (vs Lvl 0): | | | | |
| Level 1 | 0.13 (0.043)** | | | |
| Level 2 | 0.29 (0.043)*** | | | |
| Level 3 | 0.50 (0.043)*** | | | |
| Provenance (all) | | 0.17 (0.013)*** | | -0.0074 (0.012) |
| Low-trust source | | -1.05 (0.045)*** | | -0.95 (0.042)*** |
| Provenance X Low Trust | | 0.053 (0.025)* | | 0.056 (0.023)* |
| Image credibility | | | 0.63 (0.0070)*** | 0.66 (0.011)*** |
| Intercept | 3.91 (0.031)*** | 3.88 (0.026)*** | 1.38 (0.034)*** | 1.24 (0.049)*** |
| $R^2_{within}$ | 0.019*** | 0.217*** | 0.138*** | 0.296*** |
| $R^2_{between}$ | 0.006*** | 0.007*** | 0.346*** | 0.347*** |

nance information would be more beneficial to them than for sources or brands with a higher level of public trust.

We performed one-way ANOVAs to examine whether any of the effects were subject to order effects, due to being presented a specific outlet or condition first. Source trust was revealed to not be affected by either showing the either the high-trust or low-trust outlet first: $F(1, 12226) = 0.14$, $p = 0.70$. In addition, the order in which news article previews were presented neither affected source trust (all tests: $p > 0.25$), nor other evaluative aspects.

**Demographic Factors** We explored interaction effects between provenance labels and a selection of demographic information on trust, but observed no significant interaction effects. To do so, we again performed random-effects linear regression analyses, predicting source trust. For a participant's age, we observed a negative main effect on source trust (Main effect: $\beta = -0.014$, $p < 0.001$), but did not find a significant interaction effect between provenance and age ($p = 0.533$). Similarly, we observed slightly lower levels of trust among women ($\beta = -0.13$, $p = 0.033$), and found that participants with a more right-leaning political orientation were less likely to trust news sources ($\beta = -0.12$, $p < 0.001$), but observed no interactions effects with the provenance labels used (both $p > 0.3$). Finally, general media trust had a strong positive effect on source trust ($\beta = 0.55$, $p < 0.001$, $R^2_{overall} = 0.246$), but again we did not observe an interaction effect with provenance on trust ($p = 0.210$). This showed that the effects of provenance labels on source trust did not depend on demographic factors.

## Image Credibility and Transparency

Next, we examined user perceptions towards our image content. This addressed media content level effects, which were expected to be determinants of brand level effects (i.e., source trust; cf. (Strömbäck et al. 2020)). We again used multilevel linear regression models clustered at the user level. Perceived transparency increased significantly with more detailed provenance information, both when treated as an ordinal variable ($\beta = 0.72$, $p < 0.001$) and when comparing each provenance level against the no-CC baseline (Level 1: $\beta = 0.37$; Level 2: $\beta = 1.51$; Level 3: $\beta = 2.01$, all $p < 0.001$; see Models 2.1-2.2 in Table 2). These results indicate that more detailed provenance information effectively enhances transparency perceptions.


Figure 4: Effect of provenance labels on perceived image credibility for users familiar with a source or media brand versus unfamiliar users. Mean image credibility ratings (dots) and ±1 standard error bars are plotted for four labeling conditions: No CC, Logo Only, Source CC, and Full CC, separately for participants who reported using the news source ("Familiar User", red) and those who did not ("Unfamiliar User", blue). Image credibility was measured at the news source level based on three items ("The images are [fair]/[unbiased]/[accurate] when covering economy and finance topics"), using 7-point Likert scales.

The perceived credibility of images used by a specific source also increased with more detailed provenance in-

| | Model 2.1<br>$\beta$ (S.E.) | Model 2.2<br>$\beta$ (S.E.) | Model 2.3<br>$\beta$ (S.E.) |
|---|---|---|---|
| Provenance (vs Lvl 0): | | | |
| Level 1 | 0.37 (0.017)$^{***}$ | | |
| Level 2 | 1.51 (0.017)$^{***}$ | | |
| Level 3 | 2.01 (0.017)$^{***}$ | | |
| Provenance (all) | | 0.72 (0.0055)$^{***}$ | 0.74 (0.0058)$^{***}$ |
| User Familiarity | | | -0.078 (0.021)$^{***}$ |
| Provenance X Familiarity | | | 0.14 (0.012)$^{***}$ |
| Intercept | 3.12 (0.016)$^{***}$ | 3.02 (0.015)$^{***}$ | 3.00 (0.015)$^{***}$ |
| $R^2_{within}$ | $0.359^{***}$ | $0.346^{***}$ | $0.350^{***}$ |
| $R^2_{between}$ | $0.189^{***}$ | $0.179^{***}$ | $0.183^{***}$ |

Table 2: Three multilevel linear regression models predicting *perceived image transparency*, based on the extent to which provenance information was shown. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

formation (see Figure 4). Two random-effects regression models indicated significant positive effects for both ordinal provenance ($\beta = 0.28$, $p < 0.001$) and individual provenance levels relative to baseline (Level 1: $\beta = 0.087$, $p < 0.01$; Level 2: $\beta = 0.54$, $p < 0.001$; Level 3: $\beta = 0.77$, $p < 0.001$; Table 3). Thus, providing detailed provenance information enhances perceptions of image credibility, especially at higher detail levels.

**User Familiarity with a Source** We also analyzed the effects on perceptions of image credibility and transparency, considering whether they were familiar users of a specific news source or reported not to use one (i.e., unfamiliar). The main effect of familiarity and the interaction between familiarity and provenance were added to the models of transparency and credibility.

For perceived image transparency, a significant interaction was found between provenance information and user familiarity with the source (cf. Model 2.3, Table 2). Although 'familiar users' exhibited slightly lower overall transparency perceptions ($\beta = -0.078$, $p < 0.001$), the interaction indicated a stronger positive effect of provenance information among current users ($\beta = 0.14$, $p < 0.001$). Hence, provenance details particularly enhanced transparency perceptions among current readers of a given news platform, suggesting that implemented C2PA provenance labels could also benefit a platform's current readership.

For perceived image credibility, we also observed a significant interaction between provenance information and user familiarity (cf. Model 3.3 in Table 3). Here, both provenance ($\beta = 0.29$) and familiarity ($\beta = 0.25$; both $p < 0.001$) positively affected image credibility. In addition, we observed a positive, but much weaker, interaction effect between provenance and familiarity ($\beta = 0.039$, $p < 0.05$), indicating that provenance labels have a stronger effect on credibility for users who are regular users of a media brand. This suggested that familiarity was a relatively important determinant of image credibility perceptions, in addition to provenance labels.

## Mediation Analysis

Lastly, we explored whether image credibility perceptions mediated the relationship between provenance information and source trust. Employing the regression mediation method recommended by Baron and Kenny (1986); Zhao, Lynch Jr, and Chen (2010), we observed that provenance information significantly predicted image credibility (path a; Model 3.2 in Table 3), which in turn also significantly predicted source trust (path b; $\beta = 0.63$, $p < 0.001$, see Model 1.3 in Table 1). Although source trust was significantly predicted by provenance (path c; see Model 1.2 in Table 1), we found that these provenance effects were no longer present when controlling for image credibility (path a × b × c; $p = 0.528$, see Model 1.4 in Table 1. Thus, following (Baron and Kenny 1986; Zhao, Lynch Jr, and Chen 2010), we found that image credibility fully mediated the effects of provenance details on source trust, indicating that provenance labels enhanced trust primarily through improved perceptions of image credibility. Transparency also mediated provenance effects in a similar way, although image credibility had stronger mediation effects (these results are omitted for the sake of brevity).

These findings show that providing accessible and detailed provenance information significantly contributes to media trust through improved perceptions of transparency and image credibility. This aligns with the framework of Strömbäck et al. (2020), as media content effects (i.e., image credibility) predict media brand effects (i.e., source trust).

## Discussion and Conclusion

This paper has rigorously examined the effects using C2PA technology for image provenance labels in digital news. We have demonstrated that news article images accompanied by C2PA provenance labels are perceived as more transparent and fair than images without a label. In turn, the news sources who have shown such a label are attributed higher levels of trust. This all particularly holds for labels that show specific details of image provenance, such as date of creation

| | Model 3.1<br>$\beta$ (S.E.) | Model 3.2<br>$\beta$ (S.E.) | Model 3.3<br>$\beta$ (S.E.) |
|---|---|---|---|
| Provenance (vs Lvl 0): | | | |
| Level 1 | 0.088 (0.028)$^{**}$ | | |
| Level 2 | 0.54 (0.028)$^{***}$ | | |
| Level 3 | 0.77 (0.028)$^{***}$ | | |
| Provenance (all) | | 0.28 (0.0090)$^{***}$ | 0.29 (0.0095)$^{***}$ |
| User Familiarity | | | 0.25 (0.035)$^{***}$ |
| Provenance X Familiarity | | | 0.039 (0.019)$^{*}$ |
| Intercept | 4.06 (0.022)$^{***}$ | 3.99 (0.019)$^{***}$ | 4.03 (0.019)$^{***}$ |
| $R^2_{within}$ | $0.126^{***}$ | $0.118^{***}$ | $0.130^{***}$ |
| $R^2_{between}$ | $0.0251^{***}$ | $0.0239^{***}$ | $0.0413^{***}$ |

Table 3: Three multilevel linear regression models predicting *perceived image credibility*. Provenance represents four level of Content Credentials (CC): 'No CC' (lvl 0), 'Logo Only' (lvl 1), 'Source CC' (lvl 2), 'Full CC' (lvl 3). $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

and location.

We offer three primary contributions. First, we address the existing gap in the peer-reviewed literature by systematically evaluating the effects of provenance labels in a controlled experimental setting. Previous pilot studies conducted by BBC Research & Development (Monday and Strappelli 2024) lack peer-review validation and were focused on a single platform, while the only other peer-reviewed study in this domain examines provenance labels exclusively within social media platforms (Feng et al. 2023). Second, we go beyond these preliminary findings by investigating effects across multiple platforms, including both low-trust and high-trust media sources. Our research indicates that even established high-trust sources, such as Norway's NRK, can further enhance user trust through the implementation of C2PA provenance labels. At the same time, our findings highlight the potential for significant trust-building effects among lower-trust news platforms. Third, we have used a large representative sample of participants (n = 6,114) covering three different countries with distinct media landscapes (USA, UK, Norway), ensuring that our results provide robust evidence that can generalize broadly across 'western' media contexts. In doing so, we have observed some nuanced differences between countries, noting that absolute differences in source trust are due to a brand selection effect (e.g., The Sun (UK) had a comparatively low trust rating), but also differences in trust in the media landscape (Strömbäck et al. 2020; Tsfati and Ariely 2014).

An essential finding is the relationship between the amount of provenance detail provided and the degree of source trust fostered. Labels with more detailed provenance metadata significantly enhance perceived transparency and image credibility, subsequently boosting overall trust toward the news source. However, this impact on source trust appears to be a second-order effect, as indicated by relatively smaller regression coefficients and lower $R^2$ values. Thus, provenance labels primarily foster source trust when users clearly perceive detailed provenance information. This aligns with journalism trust frameworks (Strömbäck et al. 2020), highlighting transparency and image credibility as critical to building trust (Gaziano and McGrath 1986; Meyer 1988). Furthermore, our study introduces a technological approach that effectively "opens the black box" of image credentials, analogous to transparency-based trust interventions commonly explored in the domains of explainable AI and human-computer interaction (Martel and Rand 2024; Tsfati and Cappella 2003).

Our methodology extends beyond previous work (Monday and Strappelli 2024; Feng et al. 2023) by considering sources with varying baseline trust levels (Newman et al. 2024). Specifically, we find that provenance information has a greater impact on trust for news sources initially perceived as less trustworthy, likely due to a larger margin for trust improvement compared to highly reputable sources like BBC News. Larger gains in source trust are possible for source with relatively low levels of public trust. This would be an example of how individual media brands could build trust in the face of mistrust (Strömbäck et al. 2020), by focusing on improving their media content to benefit their brand (Fisher 2016; Fisher et al. 2021).

Our findings on source trust should be contextualized as a media brand effect, determined by adaptations in media content (Strömbäck et al. 2020). We show that emphasizing the veracity of media content through provenance labels benefits the credibility of media content, which in turn also supports the trustworthiness of the source (Fisher 2016). This type of trust is rooted in information theory (Strömbäck et al. 2020), for which it is postulated that the credibility of the information presented can be increased by reducing its ambiguity, based on subdimensions such as fairness, impartiality and accuracy (Gaziano and McGrath 1986; Meyer 1988; West 1994). Therefore, additional media content interventions should also focus on supporting these dimensions, supporting source or brand trust by enhancing media content

trust (Strömbäck et al. 2020).

In addition, we have observed differential effects based on user characteristics and demographics. Most notable, we considered their existing familiarity and engagement with a news source. Provenance labels had stronger effects on image perceptions among individuals who regularly consume content from the evaluated sources, even though this effect was comparatively small. This indicates that the preexisting trust levels do not limit potential gains in image perceptions and, in turn, source trust. Moreover, there is still an opportunity to enhance the trust levels of new or infrequent users through provenance labeling. This finding does contrast with the preliminary research of the BBC (Monday and Strappelli 2024), which showed that trust gains are larger among non-BBC users. In contrast, the effects of provenance labels do not seem to be moderated by demographics factors, such as age, gender, political orientation, and general media trust, which is arguably surprising. For example, we did not find evidence that younger people consume media content in a significantly different way (Galan et al. 2019), beyond a small main effect of age on source trust. One reason could be that our provenance media content is sufficiently understandable across different demographic groups, but this could be examined further in a follow-up study.

## Limitations

Our study also has certain limitations. For instance, we explored mediation between provenance, image credibility, and trust using Baron and Kenny's regression-based approach (Baron and Kenny 1986), despite critiques regarding its simplicity (cf. Zhao, Lynch Jr, and Chen (2010)). This choice was motivated by the complexity and nested nature of our dataset, deeming this mediation analysis the most feasible option. Nevertheless, the full mediation effect we observed aligns well with contemporary guidelines on regression-based mediation analysis (Zhao, Lynch Jr, and Chen 2010).

Another limitation relates to our experimental design. The controlled environment presented participants with partial news content, without enabling natural self-selection behaviors typically found in actual news consumption. While our findings robustly highlight the positive impacts of provenance labels, real-world applicability could be further confirmed through studies involving naturalistic settings.

Moreover, the explanation for differences between high-trust and low-trust sources could be confounded by the nature of the sources. NRK (Norway) and BBC (UK) are both public broadcaster, which may benefit from an additional increase in trust compared to commercial sources (cf. (Newman et al. 2024)). However, since we have observed similar differences between high-trust and low-trust sources in the USA, where both sources are commercial, we expect this to not have played a significant role in our results.

## Future Work

Consequently, we recommend future research to evaluate provenance labels on active news websites. Given that our experimental setup prominently featured the provenance label and image, this likely amplified their perceived effectiveness. News organizations involved in early C2PA pilot studies have tended toward smaller, less detailed labels to reduce visual clutter (Monday and Strappelli 2024). Therefore, future work could explore interactive provenance labels. For example, labels could initially display minimal information (our Level 1), allowing interested users to access detailed provenance metadata (Levels 2 or 3) through interactive elements such as drop-down menus triggered by clicking the Content Credentials icon. Regardless of the specific label design evaluated, we believe provenance labeling will become a significant factor in enhancing user trust within digital news ecosystems.

## Acknowledgment

This research was supported by the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through the Centre for Research-based Innovation scheme, project number 309339 and Project Reynir funded by Agenda Vestlandet.

## References

Astier, H.; and Avagnina, G. 2024. Haiti violence: Haiti gangs demand PM resign after mass jailbreak. https://www.bbc.com/news/world-latin-america-68462851. Accessed: 2025-04-25.

Bansal, G.; Nawal, A.; Chamola, V.; and Herencsar, N. 2024. Revolutionizing Visuals: The Role of Generative AI in Modern Image Generation. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(11).

Baron, R. M.; and Kenny, D. A. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6): 1173–1182.

Brenan, M. 2023. Gallup Poll Social Series: Governance - Trust in Media Survey. https://news.gallup.com/poll/512861/media-confidence-matches-2016-record-low.aspx. Accessed: 2025-04-25.

Cavaciuti-Wishart, E.; Heading, S.; Kohler, K.; and Zahidi, S. 2024. Global Risks Report 2024. Insight report, World Economic Forum.

Chan-Olmsted, S.; and Kim, J. H. 2023. Exploring the dimensions of media brand trust: a contemporary integrative approach. *Journal of Media Business Studies*, 20(1): 109–135.

Coalition for Content Provenance and Authenticity. 2025. C2PA: Coalition for Content Provenance and Authenticity. https://c2pa.org. Accessed: 2025-04-25.

Coleman, J. S. 1998. *Foundations of social theory*. Cambridge, MA: The Belknap Press. ISBN 9780674312265.

Coleman, S. 2012. Believing the news: From sinking trust to atrophied efficacy. *European Journal of Communication*, 27(1): 35–45.

Dunaway, Johanna; and Searles, Kathleen. 2023. *News and Democratic Citizens in the Mobile Era*. New York: Oxford University Press.

Egelhofer, J. L.; and Lecheler, S. 2019. Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the international communication association*, 43(2): 97–116.

Faktisk.no. 2021. Nei, somaliere kommer ikke foran i vaksinekøen. https://www.faktisk.no/artikler/jp5zx/neisomaliere-kommer-ikke-foran-i-vaksinekoen. Accessed: 2025-09-12.

Feng, K. J. K.; Ritchie, N.; Blumenthal, P.; Parsons, A.; and Zhang, A. X. 2023. Examining the Impact of Provenance-Enabled Media on Trust and Accuracy Perceptions. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Fisher, C. 2016. The Trouble with 'Trust' in News Media. *Communication Research and Practice*, 2(4): 451–465.

Fisher, C.; Flew, T.; Park, S.; Lee, J. Y.; and Dulleck, U. 2021. Improving trust in news: Audience solutions. *Journalism Practice*, 15(10): 1497–1515.

Galan, L.; Osserman, J.; Parker, T.; and Taylor, M. 2019. How young people consume news and the implications for mainstream media. Report, Reuters Institute.

Gardell, M. 2014. Crusader Dreams: Oslo 22/7, Islamophobia, and the Quest for a Monocultural Europe. *Terrorism and Political Violence*, 26(1): 129–155.

Gaziano, C.; and McGrath, K. 1986. Measuring the Concept of Credibility. *Journalism Quarterly*, 63(3): 451–462.

Halford, C. 2024. Mark the good stuff: Content provenance and the fight against disinformation. https://www.bbc.co.uk/rd/blog/2024-03-c2pa-verification-news-journalism-credentials. Accessed: 2025-04-25.

Hanitzsch, T.; Dalen, A. V.; and Steindl, N. 2018. Caught in the Nexus: A Comparative and Longitudinal Analysis of Public Trust in the Press. *The International Journal of Press/Politics*, 23(1): 3–23.

Helberger, N. 2019. On the democratic role of news recommenders. *Digital Journalism*, 7(8): 993–1012.

Holbert, R. L. 2005. Back to basics: Revisiting, resolving, and expanding some of the fundamental issues of political communication research. *Political Communication*, 22(4): 511–514.

Hwang, J.; and Oh, S. 2023. A Brief Survey of Watermarks in Generative AI. In *2023 14th International Conference on Information and Communication Technology Convergence (ICTC)*, 1157–1160.

Ipsos. 2023. Survey on the Impact of Online Disinformation and Hate Speech. Study, UNESCO and Ipsos.

Kovach, B.; and Rosenstiel, T. 2021. *The Elements of Journalism: What Newspeople Should Know and the Public Should Expect*. New York, NY: Crown, 4 edition. ISBN 0593239350.

Langguth, J.; Pogorelov, K.; Brenner, S.; Filkuková, P.; and Schroeder, D. T. 2021. Don't trust your eyes: image manipulation in the age of DeepFakes. *Frontiers in Communication*, 6.

Liu, B. 2021. In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of computer-mediated communication*, 26(6): 384–402.

Marcus, M. 2024. Embedding transparency, enhancing trust. https://www.bbc.co.uk/rdnewslabs/news/content-credentials/. Accessed: 2025-04-25.

Martel, C.; and Rand, D. G. 2024. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nature Human Behaviour*, 8(10): 1957–1967.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An integrative model of organizational trust. *Academy of management review*, 20(3): 709–734.

Meyer, P. 1988. Defining and Measuring Credibility of Newspapers: Developing an Index. *Journalism Quarterly*, 65(3): 567–574.

Monday, L.; and Strappelli, L. 2024. Does Provenance Build Trust? https://www.bbc.co.uk/rdnewslabs/news/does-provenance-build-trust. Accessed: 2025-04-25.

Newman, N.; Fletcher, R., Robertson; T., C.; Arguedas, A. R.; and Nielsen, R. K. 2024. Digital News Report 2024. Research report, Reuters Institute for the Study of Journalism.

Nsoesie, E. O.; and Ghassemi, M. 2024. Using labels to limit AI misuse in health. *Nature Computational Science*, 4(9): 638–640.

Pennycook, G.; and Rand, D. G. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7): 2521–2526.

Peterson, E.; and Allamong, M. B. 2022. The Influence of Unknown Media on Public Opinion: Evidence from Local and Foreign News Sources. *American Political Science Review*, 116(2): 719–733.

Pfänder, J.; and Altay, S. 2025. Spotting false news and doubting true news: a systematic review and meta-analysis of news judgements. *Nature Human Behaviour*.

Silverman, C. 2020. Verification handbook for disinformation and media manipulation. https://datajournalism.com/read/handbook/verification-3.

Strömbäck, J.; Tsfati, Y.; Boomgaarden, H.; Damstra, A.; Lindgren, E.; Vliegenthart, R.; and Lindholm, T. 2020. News Media Trust and Its Impact on Media Use: Toward a Framework for Future Research. *Annals of the International Communication Association*, 44(2): 139–156.

Tsfati, Y.; and Ariely, G. 2014. Individual and contextual correlates of trust in media across 44 countries. *Communication Research*, 41(6): 760–782.

Tsfati, Y.; and Cappella, J. N. 2003. Do People Watch what they Do Not Trust?: Exploring the Association between News Media Skepticism and Exposure. *Communication Research*, 30(5): 504–529.

West, M. 1994. Validating a Scale for the Measurement of Credibility: A Covariance Structure Modeling Approach. *Journalism Quarterly*, 71(1).

Wittenberg, C.; Epstein, Z.; Berinsky, A. J.; and Rand, D. G. 2024. Labeling AI-generated content: promises, perils, and future directions. *An MIT Exploration of Generative AI*.

Zerilli, J.; Bhatt, U.; and Weller, A. 2022. How transparency modulates trust in artificial intelligence. *Patterns*, 3(4).

Zhao, X.; Lynch Jr, J. G.; and Chen, Q. 2010. Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of consumer research*, 37(2): 197–206.