The Transparency Dilemma: An Experiment on How AI Disclosures Affect Credibility Perceptions and Engagement Across Topics

Sophie Morosoli[1], Emma van der Goot[1], Valeria Resendez[2], Claes de Vreese[1], Natali Helberger[1]

[1]Unitersity of Amsterdam
[2]University of Twente

s.v.morosoli@uva.nl, e.s.vandergoot@uva.nl, v.d.c.resendez@utwente.nl, c.h.devreese@uva.nl, n.helberger@uva.nl

Abstract

The media sector's credibility has been under significant scrutiny due to a rise in misinformation and the advent of generative AI (Artificial Intelligence) technologies, which pose a further threat to its credibility. Amid these challenges, transparency about the use of AI has been championed as a key solution to restore and promote credibility. Research reveals mixed reactions regarding the transparency of AI-generated content. While some studies indicate that AI-written news might be perceived as more credible than its human-produced counterparts, others find that content labelled as AI impacts credibility negatively. This study clarifies the impact of AI labels on individuals' credibility perceptions by examining the influence of different news topics. Does the usage of AI transparency labels invoke more concern about news credibility in the context of political news compared to non-political news? The effectiveness of transparency cues may be different when it concerns more serious issues, such as political news, versus less important non-political news, such as culture. We conducted a 2x2 survey experiment (N= 207) to investigate the impact of AI disclosures in the context of political news on individuals' perceptions of source credibility, perceived manipulation, and sharing intentions. Overall, AI as a news source is considered less credible, no matter the topic, yet the AI label does not increase feelings of manipulation. When it comes to sharing intention, this is negatively affected by the AI label, but only in the case of political news. These findings can help news organisations and policymakers to develop meaningful transparency labels.

Introduction

After initial excitement and hype around generative artificial intelligence (AI) and its many possibilities for media and journalism, the actual implementation of AI into journalistic products and services is riddled with concerns and [1]open ethical questions of how to do so responsibly (Diakopoulos et al. 2024) and in a way that does not further undermine the credibility of the media (Kreps and Kriner 2023; Kieslich, Diakopoulos, and Helberger 2024). One important factor in this context is transparency. Transparency is frequently mentioned as a remedy for the public's diminished trust in the media in general (Koliska 2022) and now, as the usage of algorithms and generative AI in the newsroom becomes more prevalent, transparency is also seen as a holy grail to maintain or even increase trust in the way the media embraces the technology (Diakopoulos and Koliska 2017). Even though the discussion around source labelling and perceived source credibility is not new in the field of communication science, technological and macro-level conditions have severely changed. For instance, technologically, generative AI offers unprecedented opportunities and challenges for news organisations, which are yet to be thoroughly investigated. On an organisational level, numerous professional journalistic guidelines stress the importance of transparency towards the audience about the use of AI(de-Lima-Santos and Ceron 2021; Diakopoulos et al. 2024). In a similar way, regulatory frameworks such as the recently adopted European AI Act position transparency as the frontline defence against misleading uses of synthetic content. These developments underline the importance of revisiting the dynamics behind transparency labels and source credibility.

In theory, transparency about generative AI can have different functions: informing users that and how generative AI is being used, instilling trust in the responsible use of generative AI, protecting users against manipulation, empowering them to identify synthetic content and act responsibly (e.g., by not sharing synthetic content), or simply being honest and avoiding that the audience feels manipulated (Piasecki

[1] Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

et al., 2024). In practice, research is still inconclusive on whether and how transparency labels work (Hofeditz et al. 2021; Graefe and Bohlken 2020), or what kind of information users need to be able to make informed decisions. It is possible that transparency will remain ineffective, or worse, that it will backfire, reducing rather than increasing trust in the responsible use of generative AI in the media, and reducing the credibility of AI generated news. While Hofeditz et al. (2021) found no evidence that transparency labels - text generated by AI - influenced the perceived credibility of news, other evidence suggests that disclosures backfire because individuals do not understand or misunderstand what the labels mean (Zier and Diakopoulos 2024). This, in turn, could further undermine trust in the media.

Other underlying factors may explain why transparency labels backfire in some cases and not in others. However, little is known about which such factors could influence the effect of transparency labels on individuals. This study contributes to this emerging strand of research by investigating the role of different news issues when exposing individuals to disclosures of AI generated news content. Previous survey research observed that people tend to prefer journalists over AI in political news, but mind it less when AI is used to produce sports news (Sun, Hu, and Wu 2022). We therefore measure how individuals react to political news labelled as generated by AI compared to non-political news. Specifically, we explore how transparency labels across different topics affect citizens' trust in the news source and feelings of manipulation. We expect that when an issue is deemed more important or impactful, such as in the case of political news, individuals may be less accepting of AI's involvement in writing an article and not trust AI as a competent and credible news source. However, when it comes to soft-news topics such as sports or culture, individuals might not have a problem if AI is used to create the article because they perceive it as a less sensitive topic (Sun, Hu, and Wu 2022). Hence, we believe that the topic has a moderating role when measuring the effects of transparency labels on individuals.

Another important aspect that might be impeded by using AI in journalism is citizens' engagement with information, which is an important part of public opinion formation and democratic processes (Habermas 1990). Hence, the use of AI might trigger severe news avoidance and entrenched distrust in the media. However, while multiple studies have focused on trust and credibility perceptions of AI generated news (e.g., Hofeditz et al. 2021; Tandoc, Lim, and Ling 2020), only little research has focused on the follow-up behaviour after individuals have been exposed to labelled news content (see Altay and Gilardi 2023). We fill this gap by measuring how labels affect citizens' news sharing behaviour. Yet again, to what extent citizens engage with news that is generated by AI may depend on different news topics. Therefore, we argue that it is essential to deepen and advance our understanding of how different topics and transparency labels shape not only judgments but also behaviour among individuals.

To do so, we conducted a 2 (human journalist vs AI label) by 2 (political vs. non-political topic) survey experiment (N=207). The results revealed that simple AI labelling can indeed backfire as users will receive the so-labelled news less credible – irrespective of its actual content or the issue. At the same time, not labelling AI generated content risks users feeling manipulated, which can eventually negatively affect trust. These findings illustrate the critical importance of moving beyond simple transparency labels and finding more meaningful ways to convey whether or not AI-generated or assisted content is trustworthy. This could help strengthen the relationship individuals have with news, increase trust in information and combat news avoidance.

Conceptual Framework

AI transparency labels in journalism: Aversion or appreciation

In recent years, research has focused on the risks and consequences of integrating algorithms into various journalistic practices. On one hand, algorithms represent technological innovations that help users access content more easily, generate content efficiently, gather information across platforms, among other benefits (Monzer et al. 2020). On the other hand, there are concerns that algorithms can limit diversity and lead to discrimination (Helberger, Karppinen, and D'acunto 2018). In the newsroom, AI can serve news organisations to write, translate, and summarise information more efficiently (Stray 2019; de-Lima-Santos and Ceron 2021). News writing, in particular, is a process where AI plays a critical role. However, as AI continues to evolve, challenges related to its use in journalism are becoming more apparent. For instance, Gherheș and colleagues (2025) highlight that AI-generated headlines, such as clickbait headlines, can often be indistinguishable from those written by humans, and that can affect the credibility of the source.

The growing use of AI in journalism raises additional concerns, including the absence of traditional journalistic values in AI-generated content, the potential for generating fake news and disinformation, and the need for more transparency through labels such as "written by humans" or "written by AI" (Van Dalen 2024). Although transparency through labels such "written by AI" might seem like a solution to building trust in journalistic content, it also brings new challenges for journalism. Thus, there is a growing call for more research on the circumstances and conditions under which transparency information may benefit or harm journalists (R. Wang and Ophir 2024; Diakopoulos and Koliska

2017). Particularly, since such disclosures can shape audience perceptions in different ways, influencing whether they trust or reject AI-generated content.

One possible reaction is algorithmic aversion, where individuals favour human decision-making over algorithmic tools, even when the latter demonstrate superior performance (Dietvorst, Simmons, and Massey 2015). This aversion may stem from a general scepticism towards technological developments (O'Shaughnessy et al. 2023). On the contrary, there is a growing body of research suggesting the rise of algorithmic appreciation, with users increasingly favouring AI-generated content over human-produced outputs (e.g., Logg, Minson, and Moore 2019). This shift can be attributed to the perception that AI, void of human emotions, desires, and biases, produces more impartial, accurate, and objective content (e.g., Longoni et al. 2022; Sundar 2008). In line with this perspective, Chung et al. (2023) explained that content labelled as AI can decrease motivated reasoning as individuals might overestimate the accuracy and neutrality of artificial intelligence content. To explain why individuals either resist or embrace algorithmic decision-making a possible explanation is the usage of heuristics (e.g., Logg, Minson, and Moore 2019).

Transparency labels as a cue: heuristics

An important concept explaining the user's evaluation of AI disclosures is the usage of heuristics, as disclosing the usage of AI can serve as a "symbolic cue" for users (Sundar and Kim 2019). Heuristics can be understood 'as simple procedure that helps find adequate, though often imperfect, answers to difficult questions' (Kahneman 2011). In the context of AI, one common heuristic are machine heuristics. Machine heuristics refers to "mental shortcut wherein we attribute machine characteristics or machine-like operation when making judgments about the outcome of an interaction" (Sundar and Kim 2019, 2). Depending on the appropriateness of applying these machine attributes, this shortcut can have both positive and negative effects (Yang and Sundar 2024). For example, an individual may have a positive perception of AI (i.e., algorithmic appreciation) if they associate algorithms with impartiality, but this could also lead to overreliance or uncritical acceptance of algorithmic outputs, especially in situations that require careful evaluation.

Importantly, these types of perceptions (i.e., algorithmic aversion – appreciation) can influence the credibility of the source of a piece of information (Sundar and Nass 2001). More generally, "credibility is classically ascertained by considering the source of information. If the attributed source of a piece of information is a credible person or organisation, then, according to conventional wisdom, that information is probably reliable" (Sundar 2008, 73). Source credibility has been linked to several factors, including trust, experience, honesty and impartiality (McGinnies and Ward 1980). Thus, source credibility can be measured through multiple descriptions of the source: whether the source can be considered accurate, believable, trustworthy, biased, unfair, and comprehensive (Tandoc, Lim, and Ling 2020). In this study, we follow this rationale and measure source credibility by considering these different elements. In a way, the communication of the source provides a cue for users to understand what produces the information given.

Existing evidence is mixed. For instance, there are studies that did not find that transparency about the source influenced trust in AI-written articles or human-written articles (Hofeditz et al. 2021). On the contrary, Graefe et al. (2018) suggest that individuals exposed to news labelled as generated by a machine perceived the news as more credible and with higher journalistic expertise than human-written news. The authors suggest that AI-written news can be perceived positively due to machine heuristics, where AI-generated text can be perceived without bias. However, other scholars suggest that news labelled as written by AI is perceived as less credible than content written by human journalists. It is argued that individuals have a general tendency to reject algorithms, and thus AI-generated decisions or news-making, even if they know that algorithms can outperform humans (Dietvorst, Simmons, and Massey 2015; Mahmud et al. 2022). It is argued that individuals may be concerned about AI since it tends to operate as a 'black box' and, therefore, lacks important journalistic values such as transparency or accountability (Komatsu et al. 2020). There is also empirical evidence to support this claim. Recent studies indicate that political news written by AI, and disclosed as such, is perceived as less credible than human-written news(Hong, Chang, and Tewksbury 2024; S. Wang and Huang 2024). Wang and Huang (2024) argue that AI-written news can lead citizens to form negative evaluations of the information due to prior negative experiences with the technology. Under this scenario, transparency about the fact that news was written by or with the help of AI could have a negative effect on the credibility of this piece of news. Hence, transparency labels about the use of AI in the news production might also backfire and erode trust in the media instead of fostering it. Overall, we observe a trend in more recent studies, which suggests that news labelled as AI are generally evaluated more negatively. Therefore, we hypothesise:

H1a: News content labelled as generated by AI has a negative effect on perceived source credibility compared to news labelled as written by a human journalist.

AI transparency labels and the role of political issues

To disentangle the complex effects of transparency labels on individuals' perceptions of news labelled as written by AI,

it is interesting to investigate the role of the specific news topics covered. Different topics may explain when news labelled as written by AI is deemed credible and when it is not. Thus, considering that some topics are viewed through a partisan lens (Chung, Moon, and Jones-Jang 2023), it is crucial to assess how the influence of transparency labels might vary across different issues. Individuals may find AI-written news less credible when covering a politicised topic, such as migration, compared to a non-political topic. In line with this, findings from Sun, Hu, and Wu (2022) suggest that 73.7% of their respondents prefer human journalists' reports on current affairs and politics, compared to 31.2% for sports coverage. Recent studies further show similar trends. For instance, Hong et al. (2024) found that citizens in the United States perceived human journalists as more credible than AI journalists when reporting on politically sensitive topics, even though AI journalists were seen as less biased. Similarly, Wang and Huang (2024) showed that news authored by machines triggered lower credibility perceptions on socio-political topics, suggesting that a negative machine heuristic is more likely to be activated when such topics are involved. This difference suggests that, in addition to the transparency label, the type of topic might impact individuals' perceptions and the acceptance of the use of AI in journalism in connection with source credibility, too. Therefore, we propose the following hypotheses:

H1b: This effect is stronger for political news issues compared to non-political news topics.

Perceived Manipulation

In an era of AI-generated (political) misinformation and deepfakes, the trust relationship between media companies and citizens is under increased pressure. In this context, individuals may not only lack trust but may perceive the media as trying to manipulate audiences. We understand manipulation as the process in which media undermine citizens' ability to consciously make decisions and take actions that align with their personal values and boundaries (Aïmeur and Sahnoune 2020). Previous literature on artificial intelligence and manipulation has mostly focused on the overall ability of artificial intelligence to manipulate individuals (i.e. Carroll et al. 2023; Stahl, Schroeder, and Rodrigues 2023). For instance, Stahl and colleagues (2023) argue from an ethical perspective that AI could be used to extort sensitive data from individuals (privacy infringements) and to manipulate voters. Hence, the authors see the manipulation of individuals through artificial intelligence linked to privacy and data protection. However, research is still scarce on how artificial intelligence's involvement in the news production process might trigger feelings of manipulation in individuals. As AI is increasingly used to generate misleading information and synthetic images, concerns about manipulation intensify (Wortel, Vanwesenbeeck, and Tomas 2024). Although AI itself is not inherently deceptive, its capabilities can amplify individuals' perception of manipulation when encountering AI-generated content, particularly when it mimics human journalism (Piasecki et al. 2024). Without clear disclosures on AI's role, distinguishing AI content from human content becomes nearly impossible, fuelling this perceived manipulation (Wortel, Vanwesenbeeck, and Tomas 2024).

To better understand the possible relationship between AI disclosures and perceived manipulation, we have to take a step back and look once again at source credibility in an online environment. We know that source credibility plays a crucial role when investigating manipulation. For instance, in a study on credibility and manipulation in online relationships, Aïmeur and Sahnoune (2020) explain that for manipulation to take place, manipulators rely on individuals' trust in the source to hinder them from making conscious, rational decisions. To put it simply, if individuals believe in a source, they are more likely to be manipulated because they are 'blinded' and are less likely to think rationally. Related to this, Araujo (2023) found that individuals who display higher levels of institutional trust perceive AI as less manipulative than individuals with lower levels of trust. Rui (2018) adds the perception of manipulation derives from the relationship between individuals and the source. The less trust in a source, the more individuals can perceive an intention of manipulation towards them. Furthermore, recent findings suggest that individuals cannot distinguish AI-generated from human-generated news content, which makes generative AI a dangerous tool for spreading more distrust in the media (Kreps, McCain, and Brundage 2022). This inability to distinguish and the possibility of broken 'blind' trust might leave individuals with feelings of manipulation, which can be detrimental to their information consumption and engagement – even if the information is labelled as written by AI. These feelings of manipulation, however, might be more pronounced when it comes to AI transparency labels linked to political news topics, such as migration, than softer news topics. We believe that if artificial intelligence is disclosed to be involved in generating news on political topics, individuals might be more suspicious because they think that AI could be deceiving and should not be used to write about these delicate issues. Furthermore, individuals might be aware of the pitfall of AI producing false or biased information, which might be considered more problematic in the context of political issues – even if AI is used by a news outlet they trust. Against this backdrop, we propose the following assumptions:

H2a: News content labelled as generated by AI has a negative effect on the feeling of manipulation compared to news labelled as written by a human journalist.
H2b: This effect is stronger for political news issues compared to non-political news topics.

Sharing AI generated Content

Sharing information online or offline can be considered an important part of (public) opinion formation and is therefore crucial for society. Individuals share and spread information they deem, among other things, trustworthy and accurate, stemming from credible sources (Sterrett et al. 2019). But what happens to individuals' willingness to share news when news is labelled as written by AI (AI as the news source)? And how much is this sharing intention dependent on the news topic? Only recently, scholarly attention has grown to better understand whether or how AI disclosures influence citizens' sharing intentions of regular news. Previous research rather investigated the effect of labelling of AI on other engagement measures (i.e., continue reading AI news) in the context of misinformation (Bashardoust, Feuerriegel, and Shrestha 2024; Epstein et al. 2023). These studies focusing on misinformation can provide some first insights. For instance, Bashardoust, Feuerriegel, and Shrestha (2024) compared the willingness to share AI-generated misinformation vs. human-generated misinformation and found evidence that both were being disseminated equally often by participants. In this case, participants were not directly informed that the piece of misinformation they read was AI generated. More recently, scholars have specifically focused on the direct influence of AI disclosures on citizens' news sharing intentions or engagement with AI generated news (Altay and Gilardi 2023). Altay and Gilardi (2023) asked participants of an online survey experiment how likely it is that they would share an online post of a news headline. In a nutshell, the results revealed that AI labels reduced participants' sharing intention (Altay and Gilardi 2023). However, the measurement of sharing intention used by Altay and Gilardi (2023) is one-dimensional and quite vague, and we believe that sharing can take on multiple forms. In this study, we not only measure sharing intention online on social media but also offline (talking about it) and the willingness to share news headers with peers. This measurement gives more detailed insights into how AI disclosures impact individuals' news consumption and follow-up behaviour. Furthermore, we believe that this concept of sharing intention will be influenced by the presented news topic. The willingness to share political news labelled as written by AI might be lower than the sharing intention of non-political issues. A possible explanation for this might be that individuals' reputation is at stake when they share news online and with peers (Altay, Hacquin, and Mercier 2022). The fear of sharing false information and the anticipated 'punishment' by their network will hinder individuals from engaging with news labelled as written by AI and the punishment might be considered bigger when it concerns political or more serious issues. We therefore hypothesise:

H3a: News content labelled as generated by AI has a negative effect on the sharing intention compared to news labelled as written by a human journalist.
H3b: This effect is stronger for political news issues compared to non-political news topics.

Method & Measurements

Design: This study investigates the impact of AI disclosures on individuals' perceptions of information. By means of a pre-registered 2 (Transparency label as headline source attribution: AI vs. Human) x 2 (News Topic: Political vs. Non-Political) between-subject survey experiment, we presented the participants with different news headlines with images which were created for the purpose of this study. One headline concerned a private donation made to the national museum (*non-political*) and one concerned an increase in immigrants in the Netherlands, especially because of the war in Ukraine (politicised). Next to the news issue, we manipulated the news source and presented the respondents with the name of a human journalist or AI as the writer of the respective article. Respondents were randomly assigned to the experimental conditions. We chose a layout that resembles a generic news website without any source indication to avoid priming.

Sample*:* In total 286 respondents participated in the experiment and were recruited by a panel company based in the Netherlands (Bilendi). The experiment was part of a larger survey (N = 1478), where 286 of the 1478 participants took part in the additional experiment after completing the survey. The polling company Bilendi recruited the sample based on country-specific census data and specific quotas on age, gender, and education. This resulted in a representative sample of the Dutch population regarding age (M = 50.46, SD = 17.38), gender (female = 52.9%, male = 47.1%), and education (lower = 22.1%, moderate = 50.2%, higher = 27.7%).

Procedure*:* Respondents accessed the survey experiment via an online link and had to give their informed consent before participating. The individuals were randomly assigned to one of the four experimental conditions and presented with a news header including an image. After that, participants answered questions measuring demographics and control variables. After reading the news headers, participants were forwarded to the post-test survey. This part included measures for the dependent variables and the manipulation checks. The average response time was 8 minutes. After completing the survey, participants were directed to the panel company where they received the incentives.

Independent Variables and Stimuli: Our independent variables are twofold: (1) we manipulated the source of a news header, and (2) we manipulated the news issue. The transparency label condition was manipulated by exposing

the participants to news headers, where the author of the article was adjusted. The label either stated that the headline was written by a human journalist or written by artificial intelligence. The labels were highlighted to make sure participants paid attention to them. The topic variation was manipulated by presenting the participants with two different topics. A non-political topic, where the national museum received a private donation and a heavily politicised topic covering the increase of immigrants, especially because of the war in Ukraine (see Figure 1). We purposely chose factual wording for the political issue, which does not imply a side of the political spectrum. In this study, we are interested in two different sources and issues and therefore refrained from designing stimuli catering to specific political attitudes. This would have introduced another factor in the design.

Figure 1. Political experimental stimulus with AI label

Dependent Variable: This study focuses on three core concepts: (1) source credibility, (2) perceived manipulation, and (3) sharing intention. Source credibility focuses on the trustworthiness and expertise of a source (Gaziano and McGrath 1986; Hanimann et al. 2023) and was measured through six items on a 7-point scale (1 = strongly disagree, 7 = strongly agree). The construct includes items such as: "I believe I can trust the source behind the article", "I believe the source of the news article is accurate", and "I believe the source of the news article is objective" (Cronbach's $\alpha = .93$, $M = 4.02$, $SD = .89$). The concept of perceived manipulation was measured through three items on a seven-point Likert scale (1 = strongly disagree, 7 = strongly agree) (Campbell 1995). Participants were presented with statements such as "I believe the source of the news article is able to persuade me" or "I believe the source of the news article is able to mislead me" (Cronbach's $\alpha = .65$, $M = 3.88$, $SD = .76$). To establish possible follow-up behaviour after being exposed to news headlines generated by AI or a human journalist, we measured sharing intention with the help of three items. Participants indicated on a 7-point Likert scale to what extent they agree with the following statements (1 = strongly agree; 7 = strongly disagree): (1) I would share the article on social media; (2) I would share the article with my friends; (3) I would talk about it with my friends and family, Cronbach's $\alpha = .83$, $M = 2.48$, $SD = 1.57$.

Controls: As we test political vs neutral news headers, it is important to control for political orientation to eliminate that this variable is the driving factor of the results. Political orientation was measured on a 10-point scale (0 = left, 10 = right) by asking the participants where they would place themselves ($M = 5.59$, $SD = .58$). A second control variable, which we measured was the acceptance of AI generated content connected to different news issues (e.g., sports, politics, scientific discoveries. We recorded the measure into a dummy variable where 0 = soft news topics (sport, arts and culture) and missing values and 1 = hard news (politics, international politics, medicine, scientific discoveries), $M = .56$, $SD = .03$.

Manipulation Check. For the manipulation check, we asked respondents whether the article they saw was written by a human We presented them with three answer categories: yes, no, and do not know. We find that the manipulation was successful ($\chi 2\ (2) = 27.22$, $p<0.001$). Participants exposed to a news headline labelled as written by a human journalist were significantly more likely to indicate that the article they just saw was written by a human (17.87%) compared to participants who received news headlines labelled as written by AI (13.51%). The same applied to the generated by AI condition: Individuals who were exposed to a news headline labelled as written by AI, correctly identified that the article was not written by a human journalist (22.71%) compared to the human condition (5.80%). The experimental conditions also differed from each other when looking at individuals who answered that they did not know who wrote the article. Individuals who were exposed to the articles written by AI were significantly more certain about their answer (14.98%) than individuals in the human journalist condition (25.12%).

Analysis. We removed straighteners and other outliers based on age limit, response time, and from each construct based on the cut-off value of +/3 SDs from the mean. This resulted in a final sample of 207 individuals. To assess our hypotheses and to answer the overarching research question, we performed a series of two-way analyses of variance (ANOVA). We performed a post hoc power analysis to test whether our sample size was sufficient. Based on prior research, we expect a small to moderate effect sizes ($f = 0.25$) (Hofeditz et al., 2021; Molina & Sundar, 2022; Ozanne et al., 2022). Results indicated the required sample size to

achieve 95% power for detecting a medium effect, at a significance criterion of $\alpha = 0.05$, was N = 203 for an ANOVA. Hence, the results confirmed that our samples are sufficient in size.

Results

To test the effect of news content labelled as generated by AI and political issues on source credibility (H1), we ran a two-way ANOVA. We observe a significant main effect of the AI label on source credibility ($F(1, 206) = 6.98$, $p = .01$, partial $\eta^2 = .035$). Thus, this indicates that the news labelled as AI-generated was perceived as less credible. There is no statistically significant interaction effect between news labelled as generated by AI and the different news issues ($F(1, 206) = 0.42$ $p = .52$). Hence, there is no difference between the news topics, instead, source credibility is negatively affected when labelled as AI regardless of the issue.

Further, we were interested in whether news content labelled as generated by AI increases the feeling of manipulation triggered by the label compared to news labelled as written by a human journalist and whether this effect is stronger for news covering a political issue (H2). The results reveal no statistically significant differences between the two experimental groups (generated by AI vs human journalist), $F(1, 206) = 0.75$, $p = .38$. The interaction effect between news labelled as generated by AI and the different news issues was not significant either ($F(1, 206) = 0.03$, $p = .96$). This finding points to the fact that an AI label might impact individuals' source credibility, but it does not reach as far as triggering feelings of manipulation.

Lastly, we measured if news content labelled as generated by AI has a negative effect on individuals' sharing intention compared to news labelled as written by a human journalist and whether this effect is stronger for political news compared to soft news (H3). The analysis shows no significant main effect ($F(1, 206) = 0.08$, $p = .78$). Thus, the AI label does not influence individuals sharing intention across both issues. In this case, however, we find a significant interaction effect between news labelled as generated by AI and the political news issue ($F(1, 206) = 4.54$, $p = .03$). The sharing intention of individuals is negatively affected by the AI label, but this is only the case for the political news issue (see Figure 2).

Discussion

Transparency is often regarded as the holy grail for maintaining or even increasing trust in the media (Diakopoulos and Koliska 2017). Yet today, research is still inconclusive on the effectiveness of transparency labels: some suggest that labelling improves the credibility of the source (Graefe et al. 2018), whereas others find that it backlashes, and people lose trust in the source (Zier and Diakopoulos 2024).

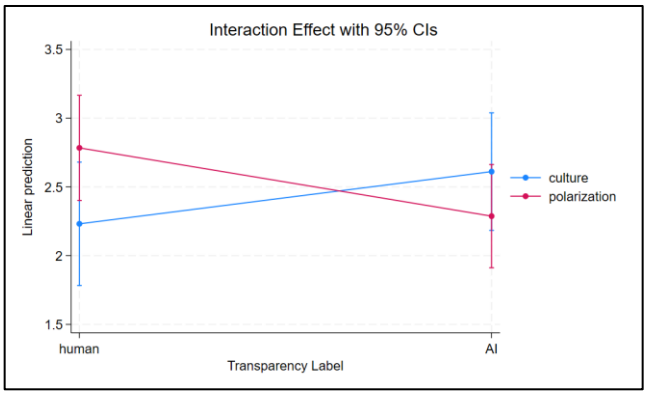

Figure 2. Interaction effect between news labelled as generated by AI and topic

To address these inconclusive findings, we investigated whether the effect of the transparency labels on individuals' perceptions and behaviour may depend on the news topic at hand. First and foremost, our results show that news content labelled as generated by AI is perceived as less credible compared to human-written news, no matter the issue. Hence, the AI label affected the source credibility negatively. While we expected that the AI label would instil less credibility in the source, specifically in the context of political news, we find that respondents do not trust AI as a news source regardless of the topic. Similar to relying on mental shortcuts such as the outlet or content, individuals may rely on the label as a heuristic cue to determine whether something is accurate (Altay and Gilardi 2023). Given media and public debates surrounding the accuracy of AI and its involvement in spreading misinformation, it is perhaps not surprising to find that people perceive AI as an unreliable new source. These findings also align with existing literature indicating that individuals are generally averse to the usage of AI in the news production process (Dietvorst, Simmons, and Massey 2015; Mahmud et al. 2022). Second, we find that individuals do not experience an increased sense of manipulation when encountering AI-labelled news content. Even with politically related news topics, citizens do not perceive AI as a threat to their ability to interpret and act on information in a self-determined manner. In other words, although AI-generated news might influence the extent to which individuals perceive a news item as credible, this does not escalate into a fear of having their autonomy compromised (Aïmeur and Sahnoune 2020). Third, our results show that AI labelling does not have an overall influence on sharing intentions, but it does discourage individuals from sharing political news content online or with their friends and family. A potential explanation for this effect could be found in the literature on reputation management (e.g. Altay et al., 2022). Individuals may hesitate to share politically charged content generated by AI due to the risk of disseminating false information, which could have negative consequences

for their reputation within their network(s). Political beliefs are often deeply ingrained and connected to one's political identity and are therefore something individuals treat with caution (Van Bavel and Pereira 2018). Especially sharing false information when it concerns these deeply held beliefs may be considered risky and individuals could fear repercussions. However, our experimental set-up did not allow us to test this mechanism, and we encourage future research to consider to what extent accuracy concerns and social reputation explain reluctance to sharing AI-labelled news.

These findings prompt further reflection. On the one hand, one could argue that scepticism towards news content, whether produced by human journalists or AI, is healthy and needed in democratic societies that face challenges like disinformation (Kyriakidou et al. 2023). This critical attitude is a key aspect of media literacy (McDougall 2019) and encourages citizens to, for instance, consult multiple sources to fact-check the information they encounter (Tsfati and Barnoy 2025). It may also be promising that we find that individuals are reluctant to share AI political news, as this means that labelling may serve as an effective strategy for limiting the spread of synthetic information. On the other hand, it also increases worries that the labelling of news as AI-generated may have a backfire effect and contribute to media distrust. The simple act of labelling something could signal to individuals that a piece of information is deceptive and dangerous content, even if this is not the case, and cause news users to reject this information. It could also contribute to increasing news avoidance (Toff and Kalogeropoulos 2020) and exposure to alternative news-outlets that embrace anti-establishment worldviews (Hameleers, Harff, and Schmuck 2023). Furthermore, leaning on the truth default theory (Levine 2014), where individuals generally assume something is real and true when they come across new information, the absence of labels could also signal to news users that certain unreliable information is truthful. A simple label stating 'generated by AI' might backfire in the sense that individuals think they cannot trust this piece of information, and that everything not labelled as such is created without the involvement of AI (which often is not the case anymore) and is inherently more reliable.

Overall, even when AI is employed responsibly, and ethical safeguards are in place to ensure the integrity of journalism the act of labelling content as AI-generated can, in some instances, result in a decline in perceived credibility and we believe it is crucial e. that media organizations therefore carefully consider how to disclose information about the role of AI in the news production process. They could do so, for instance, by clarifying the specific role AI had, that humans are still involved, and outlining how AI can be beneficial in producing reliable news content. Moreover, news organisations should be forthcoming with information about what measures have been taken to ensure that the content is still trustworthy. This way AI labels can encourage critical verification skills from citizens rather than increase their overall media distrust.

While our study provides insights into the role of AI as a news source connected to political news, our study does not come without limitations. First, this study was conducted exclusively in the Netherlands, which limits the generalizability of our findings. News consumption habits and trust in news sources can vary significantly between countries, thus the results may not fully apply elsewhere (Newman et al. 2023). Expanding future research to include participants from a wider range of political and geographic backgrounds could help determine whether these findings hold across different media and political environments. The experiment's one-shot design presents another limitation, as it restricts the ability to observe how participants' attitudes or behaviours may change over time. Media effects often accumulate with repeated exposure and future research should consider a longitudinal approach because it could offer deeper insights into how source credibility and engagement evolve. A notable portion of participants also expressed uncertainty in the manipulation check, suggesting that the manipulation may have been too subtle or that participants were not fully engaged. Refining manipulation techniques in future studies would help ensure clearer results. Additionally, this study focused solely on the news source, leaving out the important role of content in shaping audience reactions. Future research should explore the interaction between news sources and content to provide a more comprehensive understanding of how both factors influence news perception and credibility.

This study highlights the complexity surrounding the use of transparency labels in addressing public concerns about AI involvement in news production. Previous research shows that individuals are in favour of labelling AI generated news as such (Altay and Gilardi 2023), hence there seems to be a strong need for transparency when it comes to the involvement of AI in the news production process. For individuals, understanding that AI is used in journalism is crucial, yet we find that labelling news content as generated by AI could potentially intensify distrust, even when it concerns non-political issues. Hence, we believe that it is important to consider individuals' preferences and needs for transparency when deciding when to apply and how to design AI disclosures. This involvement helps to ensure that the citizens feel heard, informed, and in control, which can strengthen trust in both news sources and the use of AI in journalism. For policymakers, these findings suggest the need to carefully consider how to use labels and reconsider the assumption that AI disclosures are unnecessary when editorial oversight is in place. Therefore, policies should reflect the desire for transparency by mandating clear but careful disclosure of AI involvement in news production at all times, even when editorial review is a fact. Transparency la-

bels should be designed in a meaningful way to avoid backfire effects. This would align with public expectations and contribute to greater credibility in both AI technologies and media institutions.

References

Aïmeur, E.; and Sahnoune, Z. 2020. Privacy, Trust, and Manipulation in Online Relationships. *Journal of Technology in Human Services* 38 (2): 159–83. https://doi.org/10.1080/15228835.2019.1610140.

Altay, S.; and Gilardi, F. 2023. 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). https://doi.org/10.1093/pnasnexus/pgae403

Altay, S.'Hacquin, A.; and Mercier, H. 2022. Why Do so Few People Share Fake News? It Hurts Their Reputation. *New Media & Society* 24(6):1303–24. https://doi.org/org/10.1177/1461444820969893.

Araujo, T.; Brosius, A.; Goldberg, A. C.; Möller, J.; and de Vreese, C. 2023. Humans vs. AI: the role of trust, political attitudes, and individual characteristics on perceptions about automated decision making across Europe. *International Journal of Communication* 17(28).

Bashardoust, A.; Feuerriegel, S.; and Shrestha, Y. 2024. Comparing the Willingness to Share for Human-Generated vs. AI-Generated Fake News. arXiv. http://arxiv.org/abs/2402.07395.

Carroll, M.; Chan, A.; Ashton, H.; and Krueger, D. 2023. Characterizing Manipulation from AI Systems. In Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23). Association for Computing Machinery, New York. https://doi.org/10.1145/3617694.3623226

Chung, M.; Moon, W.; and Jones-Jang, S. 2023. AI as an Apolitical Referee: Using Alternative Sources to Decrease Partisan Biases in the Processing of Fact-Checking Messages. *Digital Journalism* 12(10): 1548-1569. https://doi.org/10.1080/21670811.2023.2254820.

Diakopoulos, N.; Helberger, N; Cools, H.; Li, C.; Kung, E.; and Rinehart, A. 2024. Generative AI in Journalism: The Evolution of Newswork and Ethics in a Generative Information Ecosystem.

Diakopoulos, N.; and Koliska, M. 2017. Algorithmic Transparency in the News Media. *Digital Journalism* 5(7): 809–28. https://doi.org/10.1080/21670811.2016.1208053.

Dietvorst, B.; Simmons, J.; and Massey, C. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology* 144(1): 114–26.

Epstein, Z.; Fang, M; Alonso Arechar, A.; and Rand. D. 2023. What Label Should Be Applied to Content Produced by Generative AI? Preprint. PsyArXiv. https://doi.org/10.31234/osf.io/v4mfz.

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

Gherheș, V.; Fărcașiu, M.; Cernicova-Buca, M.; and Coman, C. 2025. AI vs. Human-Authored Headlines: Evaluating the Effectiveness, Trust, and Linguistic Features of ChatGPT-Generated Clickbait and Informative Headlines in Digital News. *Information* 16(2): 150.

Graefe, A.; and Bohlken, N. 2020. Automated Journalism: A Meta-Analysis of Readers' Perceptions of Human-Written in Comparison to Automated News. *Media and Communication* 8(3): 50–59. https://doi.org/10.17645/mac.v8i3.3019.

Graefe, A.; Haim, M.; Haarmann, B.; and Brosius, H. 2018. Readers' Perception of Computer-Generated News: Credibility, Expertise, and Readability. *Journalism* 19(5): 595–610. https://doi.org/10.1177/1464884916641269.

Habermas, J. 1991. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. MIT Press.

Hameleers, M.; Harff, D.; and Schmuck, D. 2023. The Alternative Truth Kept Hidden From Us: The Effects of Multimodal Disinformation Disseminated by Ordinary Citizens and Alternative Hyper-Partisan Media: Evidence From the US and India. *Digital Journalism*, 1–22.

Hanimann, A.; Heimann, A.; Hellmueller, L.; and Trilling, D. 2023. Believing in Credibility Measures: Reviewing Credibility Measures in Media Research from 1951 to 2018. *International Journal of Communication* 17(2022): *22*.

Helberger, N.; Karppinen, K.; and D'acunto, L. 2018. Exposure Diversity as a Design Principle for Recommender Systems. *Information, Communication & Society* 2(2): 191–207.

Hofeditz, L.; Mirbabaie, M.; Holstein, J.; and Stieglitz, S. 2021. "Do you trust an AI-journalist? A credibility analysis of news content with AI-authorship. In Proceedings ECIS 2021 Research Papers. Marrakesh.

Hong, J.; Chang H.; and Tewksbury, D. 2024. Can AI Become Walter Cronkite? Testing the Machine Heuristic, the Hostile Media Effect, and Political News Written by Artificial Intelligence. *Digital Journalism*, 1–24. https://doi.org/10.1080/21670811.2024.2323000.

Kahneman, D. 2011. *Thinking Fast, Thinking Slow*. London: Tavistock.

Kieslich, K.; Diakopoulos, N.; and Helberger, N. 2024. Anticipating Impacts: Using Large-Scale Scenario-Writing to Explore Diverse Implications of Generative AI in the News Environment. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00497-4.

Koliska, K. 2022. Trust and Journalistic Transparency Online. *Journalism Studies* 23(12): 1488–1509. https://doi.org/10.1080/1461670X.2022.2102532.

Komatsu, T.; Gutierrez Lopez, M.; Makri, S.; Porlezza, C.; Cooper, G.; MacFarlane, A.; and Missaoui, S. 2020. AI Should Embody Our Values: Investigating Journalistic Values to Inform AI Technology Design. In Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society. Tallinn Estonia: ACM. https://doi.org/10.1145/3419249.3420105.

Kreps, S.; and Kriner, D. 2023. The Potential Impact of Emerging Technologies on Democratic Representation: Evidence from a Field Experiment. *New Media & Society*. https://doi.org/10.1177/14614448231160526.

Kreps, S.; McCain, R.; and Brundage, M. 2022. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science* 9(1): 104–17. https://doi.org/10.1017/XPS.2020.37.

Kyriakidou, M.; Morani, M.; Cushion, S.; and Hughes, C. 2023. Audience Understandings of Disinformation: Navigating News Media through a Prism of Pragmatic Scepticism. *Journalism* 24(11): 2379–96. https://doi.org/10.1177/14648849221114244.

Levine, T. R. 2014. Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology* 33(4): 378–92. https://doi.org/10.1177/0261927X14535916.

Logg, J.; Minson, J.; and Moore, D. 2019. Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes* 151: 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005.

Longoni, C.; Fradkin, A.; Cian, L.; and Pennycook, G. 2022. News from Generative Artificial Intelligence Is Believed Less. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. Seoul, Republic of Korea: ACM. https://doi.org/10.1145/3531146.3533077.

Mahmud, H.; Najmul Islam, A.; Ishtiaque Ahmed, S.; and Smolander, K. 2022. What Influences Algorithmic Decision-Making? A Systematic Literature Review on Algorithm Aversion."*Technological Forecasting and Social Change* 175. https://doi.org/10.1016/j.techfore.2021.121390.

McDougall, J. 2019. Media Literacy versus Fake News: Critical Thinking, Resilience and Civic Engagement. *Medijske Studije* 10(19): 29–45. https://doi.org/10.20901/ms.10.19.2.

McGinnies, E.; and Ward, C. 1980. Better Liked than Right: Trustworthiness and Expertise as Factors in Credibility. *Personality and Social Psychology Bulletin* 6(3): 467–72.

Monzer, C.; Moeller, J.; Helberger, N.; and Eskens, S. 2020. User Perspectives on the News Personalisation Process: Agency, Trust and Utility as Building Blocks. *Digital Journalism* 8(9): 1142–62. https://doi.org/10.1080/21670811.2020.1773291.

Newman, N.; Fletcher, R.; Eddy, K.; Robertson, C.; and Kleis Nielsen, R. 2023. Reuters Institute Digital News Report 2023.

O'Shaughnessy, M.; Schiff, D.; Varshney, L.; Rozell, C; and Davenport, M. 2023. What Governs Attitudes toward Artificial Intelligence Adoption and Governance? *Science and Public Policy* 50(2): 161–76. https://doi.org/10.1093/scipol/scac056.

Piasecki, S.; Morosoli, S.; Helberger, N; and Naudts, L. 2024. AI-Generated Journalism: Do the Transparency Provisions in the AI Act Give News Readers What They Hope For? *Internet Policy Review* 13(4). https://doi.org/10.14763/2024.4.1810.

Rui, J. 2018. Source–Target Relationship and Information Specificity: Applying Warranting Theory to Online Information Credibility Assessment and Impression Formation. *Social Science Computer Review* 36(3): 331–48. https://doi.org/10.1177/0894439317717196.

Santos, M.; and Ceron, W. 2021. Artificial Intelligence in News Media: Current Perceptions and Future Outlook. *Journalism and Media* 3(1): 13–26. https://doi.org/10.3390/journalmedia3010002.

Stahl, B.; Schroeder, D.; and Rodrigues, R. 2023. *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges*. SpringerBriefs in Research and Innovation Governance. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-17040-9.

Sterrett, D.; Malato, D.; Benz, J.; Kantor, L.; Tompson, T.; Rosenstiel, T.; Sonderman, J.; and Loker, K. 2019. Who Shared It?: Deciding What News to Trust on Social Media. *Digital Journalism* 7(6): 783–801. https://doi.org/10.1080/21670811.2019.1623702.

Stray, J. 2019. Making Artificial Intelligence Work for Investigative Journalism. *Digital Journalism* 7(8): 1076–97. https://doi.org/10.1080/21670811.2019.1630289.

Sun, M.; Hu, W.; and Wu, Y. 2022. Public Perceptions and Attitudes Towards the Application of Artificial Intelligence in Journalism: From a China-Based Survey. *Journalism Practice*. https://doi.org/10.1080/17512786.2022.2055621.

Sundar, S. 2008. The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. Cambridge, MA: MacArthur Foundation Digital Media and Learning Initiative

Sundar, S.; and Kim, J. 2019. Machine Heuristic: When We Trust Computers More than Humans with Our Personal Information.In Proceedings of the 2019 CHI Conference on human factors in computing systems.

Sundar, S.; and Nass, C. 2001. Conceptualizing Sources in Online News. *Journal of Communication* 51(1): 52–72.

Tandoc, E.; Lim, D.; and Ling, R. 2020. Diffusion of Disinformation: How Social Media Users Respond to Fake News and Why. *Journalism* 2 (3): 381–98. https://doi.org/10.1177/1464884919868325.

Toff, B.; and Kalogeropoulos, A. 2020. All the News That's Fit to Ignore. *Public Opinion Quarterly* 84(S1): 366–90. https://doi.org/10.1093/poq/nfaa016.

Tsfati, Y.; and Barnoy, A. 2025. Media Cynicism, Media Skepticism and Automatic Media Trust: Explicating Their Connection with News Processing and Exposure. *Communication Research*, 00936502251327717.

Van Bavel, J.; and Pereira, A. 2018. The Partisan Brain: An Identity-Based Model of Political Belief. *Trends in Cognitive Sciences* 22(3): 213–24. https://doi.org/10.1016/j.tics.2018.01.004.

Van Dalen, A. 2024. Revisiting the Algorithms behind the Headlines. How Journalists Respond to Professional Competition of Generative AI. *Journalism Practice*, 1–18.

Wang, R.; and Ophir, Y. 2024. Behind the Black Box: The Moderating Role of the Machine Heuristic on the Effect of Transparency Information about Automated Journalism on Hostile Media Bias Perception. *Journalism*.

Wang, S.; and Huang, G. 2024. The Impact of Machine Authorship on News Audience Perceptions: A Meta-Analysis of Experimental Studies. *Communication Research* 51(7): 815–42. https://doi.org/10.1177/00936502241229794.

Wortel, C.; Vanwesenbeeck, I.; and Tomas, T. 2024. Made with Artificial Intelligence: The Effect of Artificial Intelligence Disclosures in Instagram Advertisements on Consumer Attitudes. *Emerging Media* 2(3): 547–70. https://doi.org/10.1177/27523543241292096.

Yang, H.; and Sundar,S. 2024. Machine Heuristic: Concept Explication and Development of a Measurement Scale. *Journal of Computer-Mediated Communication* 29(6).

Zier, J.; and Diakopoulos, N. 2024. Labeling AI-Generated News Content: Matching Journalist Intentions with Audience Expectations. In Proceedings of Proc. Computation + Journalism Symposium.