Security Benefits and Side Effects of Labeling AI-Generated Images

Sandra Höltervennhoff*
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Jonas Ricker*
Ruhr University Bochum
Bochum, Germany

Maike M. Raphael
Leibniz University Hannover
Hannover, Germany

Charlotte Schwedes
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Rebecca Weil
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Asja Fischer
Ruhr University Bochum
Bochum, Germany

Thorsten Holz
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Lea Schönherr
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Sascha Fahl
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Abstract

Generative artificial intelligence is developing rapidly, impacting humans' interaction with information and digital media. It is increasingly used to create deceptively realistic misinformation, so lawmakers have imposed regulations requiring the disclosure of AI-generated content. However, only little is known about whether these labels reduce the risks of AI-generated misinformation.

Our work addresses this research gap. Focusing on AI-generated images, we study the implications of labels, including the possibility of mislabeling. Assuming that simplicity, transparency, and trust are likely to impact the successful adoption of such labels, we first qualitatively explore users' opinions and expectations of AI labeling using five focus groups. Second, we conduct a pre-registered online survey with over 1 300 U.S. and EU participants to quantitatively assess the effect of AI labels on users' ability to recognize misinformation containing either human-made or AI-generated images. Our focus groups illustrate that, while participants have concerns about the practical implementation of labeling, they consider it helpful in identifying AI-generated images and avoiding deception. However, considering security benefits, our survey revealed an ambiguous picture, suggesting that users might over-rely on labels. While *inaccurate* claims supported by labeled AI-generated images were rated less credible than those with unlabeled AI-images, the belief in *accurate* claims also decreased when accompanied by a labeled AI-generated image. Moreover, we find the undesired side effect that human-made images conveying *inaccurate* claims were perceived as more credible in the presence of labels.

CCS Concepts

• **Security and privacy** → **Usability in security and privacy**; • **Information systems** → *Social networks*; • **Human-centered computing** → *Empirical studies in HCI*.

Keywords

AI-Generated Content, Misinformation, AI Labels, Social Media

*Equal contribution.

1 Introduction

Since the beginning of the AI boom [1], artificial intelligence (AI) has increasingly permeated our digital lives. While a few years ago, state-of-the-art generative AI (GenAI) methods and tools required specialized knowledge and extensive computational resources, easy-to-use tools like ChatGPT, Midjourney, ElevenLabs, and Sora enable laypeople to create highly realistic text, images, audio, and videos using natural language prompts. However, besides the countless productive and creative applications of GenAI, there is significant potential for misuse. On a number of occasions, scammers have used voice cloning and deepfakes to impersonate high-ranking staff members, tricking companies into transferring millions to the criminals [2]–[4]. Deepfakes can also harm individuals, for instance, by generating explicit content of celebrities and ordinary people without their consent [5], [6].

Another major AI-generated content (AIGC) threat is the spread of AI-generated misinformation. This became increasingly apparent during the 2024 U.S. presidential election [7]–[9], when, for instance, Donald Trump shared fabricated images that implied he was endorsed by Taylor Swift [10]. Similarly, a generated image supposedly showing an explosion near the Pentagon caused a dip in the stock market [11], highlighting the potential of AIGC to cause outrage and influence public opinion. Furthermore, the growing capabilities and ubiquity of GenAI create a "liar's dividend", making it possible to cast doubt on any content that is unpleasant or contradicts a certain narrative [12].

Given that a significant proportion of Internet users obtain at least part of their news from social media (54 % of adults in the United States [13]), the general increase in AIGC on these platforms [14] has prompted legislators worldwide to consider regulatory measures. An essential goal is the transparent disclosure of AIGC through the use of *labels*. The European Parliament passed both the Digital Services Act (DSA) and Artificial Intelligence Act (AI Act), obligating large online platforms and search engines as

well as providers of GenAI to disclose AI-generated and manipulated content. Under the presidency of Joe Biden, the U.S. government attempted to implement similar rules through an executive order, supporting the development of authentication and labeling standards [15]. However, due to the political change in political direction under President Donald Trump, it is unclear how GenAI will be regulated in the U.S. [16]. Recently, the Chinese government also published a set of rules on how AI service providers and content-sharing platforms must label AIGC [17], [18]. These legislative efforts are motivated by the expectation that AI labels will help users spot AI-generated misinformation. In contrast to existing misinformation warnings, which require manual fact-checking, several mechanisms exist to automatically detect AIGC on a large scale. This could give AI labels the potential to efficiently mitigate the threats that misleading AIGC poses to our society.

However, AI labels are primarily a transparency mechanism. They do not communicate whether the content is true or false, but only whether it was generated using AI. Therefore, research is urgently needed into the *actual* security benefits and risks of AI labeling as an instrument against misinformation.

To the best of our knowledge, we are the first to address this research gap through a mixed-methods study. First, we conducted five semi-structured focus groups to understand users' expectations, concerns, and trust in AI labels. Second, through a pre-registered online survey, we quantitatively measured how labeling affects users' belief in misinformation. As a large proportion of online misinformation involves images [19], which have been found to increase user engagement particularly on social media [20], we focus our work on the labeling of AI-generated images (AIGIs). While previous work [21] showed that AI labels can reduce users' belief in misleading AIGIs, the survey did not include other types of stimuli, providing only a restricted view on the implications of labeling. To investigate potential side effects, our stimuli set is varied in terms of image type and veracity. We study the effects of human-made and AI-generated images conveying either true or false information. Moreover, we consider the potential of mislabeling, i.e., AIGIs that are not labeled and human-made images that are wrongly labeled as AI-generated. We formulate the following three research questions:

RQ1. *What are users' opinions, expectations, and concerns about AI labeling?* Like other security mechanisms, AI labels might suffer from low adoption rates if users do not see or distrust their merit. We, therefore, explore factors that influence users' acceptance, comprehension, and trust in labels and identify potential problems that could hinder adoption.

RQ2. *How does AI labeling affect users' perception of true and false claims with and without AIGIs?* While AI labels are considered a remedy for the growing threat of deceptive AIGC, labeling may have unintended side effects on social media. We study the effects of labels on different kinds of posts and explore users' perspectives on the implications of labeling.

RQ3. *How does mislabeling interfere with the efficacy of and trust towards AI labeling?* At least within the near future, labeling mechanisms will not be without errors, resulting in unlabeled AIGIs and falsely labeled human-made images. Besides the primary security risk of unlabeled misinformation, we qualitatively and quantitatively study the broader impact of mislabeling.

2 Background

In this section, we outline existing legislation and ongoing legislative procedures regarding AI labeling, focusing on the situation in the EU and the U.S. We additionally summarize the GenAI policies of popular social media platforms as of March 2025. We provide supplementary information on technical labeling mechanisms in Appendix A.

2.1 Legal Regulation of AI-generated Content

EU. The AI Act [22] entered into force in August 2024 and establishes a comprehensive legal framework for all types of AI, aiming to protect the public against its potential risks. AI applications are categorized into four levels, from those with unacceptable risk, which are banned, to those with minimal risk, which remain unregulated. As stated in Article 50 [23], providers of GenAI are required to label synthetic content (including images, audio, videos, and text) using markings that are "effective, interoperable, robust and reliable as far as this is technically feasible". Stricter regulations are imposed on AI systems capable of creating deepfakes, which are defined as "AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful." In addition to the markings that are required for all synthetic content, providers of deepfake applications must "clearly and distinguishably disclose that the content has been artificially created or manipulated," thus, in a human-readable manner.

Another EU regulation handling the labeling of AIGC is the DSA [24], which was adopted in October 2022. It requires very large online platforms and search engines with more than 45 million monthly active users in the EU to make deepfakes "distinguishable through prominent markings when presented on their online interfaces". Moreover, platforms must offer "easy to use functionality which enables recipients of the service to indicate such information".

United States. In July 2023, several large platforms, including OpenAI, Google, and Meta, made voluntary commitments to ensure the safe use of AI, including the watermarking of AIGC [25]. The pledge was likely a reaction to previous suggestions by former President Joe Biden to regulate the use of AI [26]. It was followed by an executive order on the "safe, secure, and trustworthy development and use of artificial intelligence" [15]. Among others, it imposed transparency obligations on AI developers and ordered the National Institute of Standards and Technology (NIST) to "develop effective labeling and content provenance mechanisms", explicitly mentioning detection and watermarking. The executive order was complemented in March 2024 by a bipartisan bill requiring a "clear and conspicuous notice" to inform consumers if media was created or edited using AI [27]. However, the Trump administration revoked said executive order [16] and instructed AI researchers to prioritize performance over safety [28], making the future of AI regulation in the United States uncertain.

2.2 GenAI Policies of Social Media Platforms

Facebook/Instagram. AIGC and other forms of manipulated media are allowed on the platforms if they do not violate other policies [29]. Content created or modified with Meta's AI tools and

directly shared to Facebook or Instagram may receive a visible "Imagined with AI" watermark. Uploaded content that is found to be AI-generated based on Coalition for Content Provenance and Authenticity (C2PA) [30] or International Press Telecommunications Council (IPTC) [31] metadata will be labeled with an "AI Info" label [32]. These standards embed information on the origin and authorship into the media file and, in the case of C2PA, bind it to the content using cryptographic signatures. In addition, users are required to self-disclose the use of GenAI if they share realistic videos or audios. Failing to do so may be penalized [33].

X. X, formerly known as Twitter, does not allow "misleading media" that may cause widespread confusion or serious harm, be it synthetic, manipulated, or shared out of context [34]. Manipulated or out-of-context media that is shared in non-deceptive ways is explicitly allowed. X claims to "use [their] own technology or receive reports through partnerships with third parties" to identify synthetic or altered media but may not take actions if a reliable detection is not possible. At the time of writing, X does not employ uniform labels to mark AIGC. However, the platform's *Community Notes* system can be used to annotate manipulated media [35]. Images generated using X's *Grok* contain a visible watermark [36].

TikTok. On TikTok, AIGC or edited media depicting realistic scenes or people must be labeled using a dedicated label, caption, watermark, or sticker [37]. In contrast, no disclosure is required if content is edited without changing its core meaning. AIGC featuring minors, private figures (without their consent), or misleading claims about crises or public figures is forbidden. However, the guidelines explicitly allow the use of AI for educational purposes and the depiction of public figures in humorous settings. In May 2024, TikTok adopted the C2PA standard to add an "AI-generated" label to AIGC uploaded to the platform [38]. C2PA metadata will be embedded into content created with TikTok's own GenAI tools.

YouTube. YouTube requires content creators to self-disclose the use of altered or synthetic content [39]. This policy is not limited to GenAI but applies to content that appears realistic and is meaningfully altered or synthetically created. While for regular content the label "Altered or synthetic content" is added to the video's expanded description field, an overlay is added to sensitive content, which, e.g., relates to elections, conflicts, or natural disasters. Users who repeatedly do not disclose AIGC may have their content removed or be suspended from the YouTube Partner Program. YouTube *Shorts* created using the platform's GenAI tools are automatically labeled without any user action [39]. The platform additionally follows the complementary strategy of displaying a "Captured with a camera" disclosure in the video description if C2PA metadata confirms that the content is *authentic* [40].

LinkedIn. Content created with the help of GenAI must comply with LinkedIn's professional community policies [41], [42], which prohibit sharing false or misleading content. More specifically, users may not share misleading synthetic or manipulated media without disclosing its altered nature. Without disclosure, and if content is clearly not parody or satire, LinkedIn may remove content or limit its distribution [43]. In May 2024, LinkedIn adopted the C2PA standard and displays *Content Credentials* for images containing the respective metadata [44].

3 Methodology

We now provide details on how we conducted our focus groups and online survey. For both, we explain their procedure, recruitment, analysis, and limitations. For the survey, we also give details on the stimuli selection.

3.1 Focus Groups

Between December 2024 and February 2025, we conducted five semi-structured focus groups with three to five participants per group ($N = 18$) to foster discussions and identify strengths and problems of AI labels.

Procedure. To reach a diverse and international group of participants, we ran focus groups online using video conferencing software. At the beginning of each focus group, we informed the participants about the purpose and content and asked for consent to record the focus group. We prepared a set of initial questions (see Appendix B.2) to guide the discussion. Central topics were (A) experiences with and risks of AIGIs, (B) opinions and expectations of AI labeling, and (C) potential problems of labels and labeling mechanisms. We focused on the mechanisms *self-disclosure*, *metadata*, and *AI detection*, as they are most relevant for current implementations. Before the main study, we conducted two pilots to test our focus group guide. We occasionally adjusted existing questions or added new ones between focus groups if new perspectives came up. During the conversation, participants were shown a slide deck containing the current question and exemplary images or visualizations.

Recruitment. We recruited participants through Prolific. Participants had to live in the EU or the U.S., be over the age of 18 and have a sufficient level of English. Our study information and consent form can be found in Appendix B.1. Based on the participants' answers to a short pre-screening questionnaire, we formed groups considering age, gender, country of residence, social media usage, and their attitude toward technology interaction (ATI-S [45]) and AI (AIAS-4 [46]). In total, we conducted six focus groups. For one group, only two participants showed up for the scheduled call, who also had insufficient English skills, which is why we excluded the results from our analysis. We reached thematic saturation after the five remaining focus groups and therefore stopped recruiting. Focus groups lasted an average of 74 minutes. Participants were compensated £23.75 for an estimated duration of 90 minutes to account for unexpected events. We provide the aggregated demographics of our participants in Table 1.

Analysis. All focus groups were transcribed by a GDPR-compliant transcription service and coded using ATLAS.ti, utilizing an open coding approach. Each focus group was independently coded by two researchers, who afterward discussed the codes and agreed on a shared coding. A total of three researchers were involved in the whole process. As these discussions, including the resolving of conflicts, were crucial for forming our final codebook, we did not calculate the inter-rater reliability, which is in line with previous work [47]–[49]. The final codebook is provided in Appendix B.3. Through affinity mapping, we condensed our codebook into relevant themes that we report in Section 4.

Limitations. The ATI-S [45] and AIAS-4 [46] scores of our participants were relatively high. This may be caused by our sampling

Table 1: Demographics of our 18 focus group participants.

Gender	**N**	**Age**	**N**
Female	7	18–29	6
Male	10	30–49	10
Non-binary	1	50–69	2
Region		**ATI-S [45]**	
EU	12	$0 \leq x \leq 2$	2
U.S.	6	$2 < x \leq 4$	6
Education		$4 < x \leq 6$	10
Secondary school		**AIAS-4 [46]**	
University/College w/o degree	10	$0 \leq x \leq 2$	1
Associate degree	2	$2 < x \leq 4$	1
Bachelor's degree	4	$4 < x \leq 6$	16
Master's degree	5		

platform Prolific, as it probably attracts more tech-savvy people, who might also be more open towards new technology. Our results might suffer from biases, like self-selection, self-reporting, or a desirability bias. Moreover, group dynamics might have influenced participants' answers. To counter these effects, we tried to make participants comfortable and stressed that we will not judge any answers but are interested in their diverse perspectives. While we are convinced of our high data quality, qualitative work is not generalizable. Therefore, we will use qualifiers instead of numbers when reporting our results to not give the impression of a quantitative evaluation.

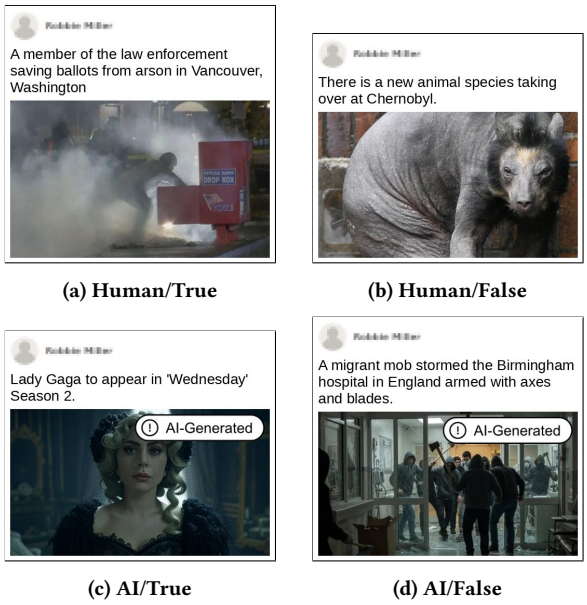
3.2 Survey

We conducted our experiment through an online survey with $N = 1\,354$ valid participants in April 2025. The pre-registration, describing our hypotheses and analysis plan, is available at osf.io/f6ztr. Before our main study, we tested the survey with colleagues and conducted three pilots on Prolific with six to 15 participants.

Procedure. Each participant was randomly assigned to one of three groups: control (C), labeling (L), and mislabeling (M) group. At the beginning, participants were informed about the purpose of the study and the use of their data. After giving their informed consent, they were instructed about the upcoming task, which was to "identify posts containing false claims that appeared on a social media platform". Participants in the control group were told that there have been AIGIs on the platform before, while participants in the treatment groups were told that the platform uses a labeling system to flag AIGIs.

In the main part of the survey, participants were shown 26 simulated social media posts, each consisting of a caption and an image. The post author's profile image and name were concealed. The stimuli varied in two dimensions: (1) whether the post's image was human-made or AI-generated (Image Type) and (2) whether the post's overall claim was true or false (Veracity). Posts were equally divided into four subsets (human/true, human/false, AI/true, AI/false; see examples in Figure 1), each containing six posts. Two additional posts served as attention checks.

Participants in the control group saw no AI labels. In the labeling group, all posts containing an AIGI were labeled as *AI-generated* in the top-right corner (see Figure 1), regardless of whether the claim was true or false. We opted for this wording, as previous work [50] has shown that people correctly associate it with AIGC. While participants in the mislabeling group also saw labels, two



(a) Human/True **(b) Human/False**

(c) AI/True **(d) AI/False**

Figure 1: Example stimuli from all four subsets.

out of the six posts in each subset were mislabeled, i.e., AIGIs were not labeled (false negative) and human-made images were labeled as AI-generated (false positive). We applied counterbalancing to ensure that each post was mislabeled equally often.

For each post, we asked participants whether they believe in the post's claim or not. Additionally, we asked participants how confident they are in their assessment. The questions were displayed after a short delay to guide the initial attention of the participants to the post. After the main part of the survey, we asked participants in the labeling and mislabeling group about their perception of the labels and their opinion on mislabeling. Finally, we collected demographics and debriefed all participants. We provide the full questionnaire in Appendix C.4.

Stimuli. Following previous studies [21], [51], [52] we sourced social media posts from popular fact-checking sites (e.g., snopes.com, factcheck.afp.com) to create realistic conditions. Topics ranged from world news and politics to lifestyle and celebrity news (see Appendix C.5). In an attempt to reduce partisan bias, we avoided stimuli that were clearly left- or right-leaning. We ensured that topics are balanced over all four subsets. For posts in the human/true, human/false, and AI/false subsets, we kept the caption as close to the original as possible, only adjusting texts that were too emotional or did not clearly purport the post's claim. We provide a list of all original and adjusted captions in Appendix C.5.

Our study design requires the AI/true subset to analyze whether AI labels simply decrease participants' belief in a claim or whether they assist them in making better judgments. Unfortunately, we were unable to find enough posts conveying true information illustrated with an AIGI "in the wild". As a remedy, we took false posts with AIGIs and adjusted the captions to make the associated claims actually true. While this somehow contradicts our goal to study real social media posts, we consider it likely that such posts will be relevant in the near future. Individual news articles are already

Table 2: Demographics of our 1354 valid survey participants. Note that not all numbers add up to the total amount, as participants had the option not to answer.

Gender	**N**	**%**	**Age**	**N**	**%**
Female	674	49.8	18–24	242	17.9
Male	660	48.7	25–34	489	36.1
Non-binary	17	1.3	35–44	276	20.4
Region			45–54	182	13.4
U.S.	672	49.6	55–64	119	8.8
EU	682	50.4	65+	45	3.3
Education			**Political views**		
10th grade or less	19	1.4	Very left	266	19.6
Secondary school	161	11.9	Left leaning	417	30.8
Trade/technical/vocational	40	3.0	Center	256	18.9
University/College w/o deg.	216	16.0	Right leaning	250	18.5
Associate degree	93	6.9	Very right	62	4.6
Bachelor's degree	469	34.6	Not interested	84	6.2
Master's degree	295	21.8			
Professional degree	22	1.6			
Doctoral degree	35	2.6			

illustrated with AIGIs taken from stock image sites [53], [54]. Moreover, first news outlets are using GenAI to compose articles, as the example of a German tabloid shows, where 11% of all articles are written by an "AI journalist" [55], [56]. Such developments indicate a beginning normalization of the use of AI to produce news content. We therefore deem the AI/true subset important to explore how AI labeling would affect the perceived credibility of such content.

Recruitment. Similar to our focus groups, participants were recruited via Prolific. We balanced our sample regarding country of residence (50% EU and 50% US participants) and gender. Moreover, we added a screener to filter for participants with an approval rate of at least 90% and more than 50 completed surveys, that had fluent English skills and used any social media platform. Participants agreed to a consent form before starting the survey, shown in Appendix C.3. Each participant was paid £2.86 for an expected duration of 16 minutes. The actual median completion time was 10 minutes. We received 1 405 completed surveys, out of which we excluded 51 participants that failed one or both attention checks. This resulted in 1 354 valid submissions for our analysis. Table 2 lists the demographics of our sample.

Analysis. If not stated otherwise, we perform all analyses according to our pre-registration. To distinguish different influences present in participants' response behavior, we follow the recommendation by Batailler *et al.* [57] to use signal detection theory (SDT) measures. According to SDT, the sensitivity measure d' indicates the ability to accurately distinguish between true and false claims. Sensitivity is calculated as $d' = z(H) - z(FA)$, where the hit rate H refers to the proportion of true judgments for true claims divided by the total number of true claims, and the false alarm rate FA refers to the proportion of true judgments for false claims divided by the total number of false claims.

SDT measure calculations do not allow for minimizing the effects introduced by differences between the used materials (i.e., images) and between study participants. Accordingly, we also determined accuracy by coding all congruently judged claims (i.e., true (false) claims judged as true (false)) with 1 and all incongruently judged claims with 0, which allowed us to use it as a dependent variable in a generalized linear mixed model (GLMM). Models included by-subject (e.g., participant) and by-item (e.g., image) random intercepts, so that we can generalize the effects beyond our specific materials and participants.

To be able to compare the three groups in our design (control, labeling, mislabeling) in line with the predicted effects, we use independent contrasts. Specifically, in our design, Helmert contrasts [58] allow for a comparison between control group vs. *both* treatment groups (i.e., first Helmert contrast), independently of the difference between the treatment groups. Additionally, with the second Helmert contrast, the two treatment groups (i.e., labeling vs. mislabeling) can be compared. To find and select the model that best fits our data, we compare our models using the Akaike information criterion (AIC). A lower AIC indicates a better model fit [59]. We provide additional information on the used software in Appendix C.2.

Limitations. To specifically investigate the effect of AI labels on security, we created an artificial study setting where we just focused on human-made and AI-generated images in the context of news. Moreover, we equally distributed our images onto all four categories (human/AI, true/false). This may have resulted in study artifacts and material effects not present in realistic social media feeds or missed interaction effects with other content. While we tried to account for diverse participants, our sample is not representative and might suffer from a self-selection bias. However, we argue that Prolific has been found to produce acceptable results in previous work [60]–[62].

3.3 Ethics Considerations

Our methodology was approved by our institution's ethical review board under the applications 24-09-1 (focus groups) and 25-01-4 (survey). Before the study, our focus group and survey participants were informed about its purpose and agreed to a consent form, also containing an option for withdrawal. The respective consent forms can be found in Appendices B.1 and C.3. We minimized the collection of personally identifiable information and pseudonymized our focus group participants' names before the analysis. We adhered to the General Data Protection Regulation (GDPR) for data collection, storage, and processing of participants' data.

To investigate the influence of AI labels on misinformation, we had to show participants misleading social media posts. However, we carefully crafted our stimuli set to not upset or frighten our participants. To foster discussions, we presented a few AIGIs in our focus groups. However, we always explained their context and disclosed false claims. While we could not fully disclose our goal to investigate AI labels when recruiting participants for our survey, we added a disclaimer that it contains misinformation that might touch on sensitive topics. Before participants submitted the survey, they were shown a debriefing page, which was also taken into account for estimating the survey duration. On this page participants were clearly informed about which posts contained misinformation and/or were accompanied by an AIGI. We also provided a link to a fact-checking article for each post. As we carefully selected our images and all shown misinformation was strictly connected to our research, we deem the risks of our study acceptable.

4 User Perception and Concerns of AI Labels

In the following, we will present participants' opinions and concerns about AI labels. While these results are primarily informed by our focus groups, we will present quantitative results from our survey where appropriate. When reporting about our focus groups, we use quantifiers to describe how many of our participants gave certain answers ("few": 2–5, "some": 6–9, "many": 10–13, "most": 14–17).

4.1 The Problem: AI-Generated Images Can Be Harmful and Hard to Spot

To explore the need for AI labels, we initially asked participants which problems they associate with AIGIs. The most prominent concern was that AIGIs can convey misinformation, with the danger of skewing the public opinion, e.g., regarding political topics. Participants also talked about the risk of forged evidence to support false accusations, with a few participants even questioning whether images can nowadays serve as proof at all:

> "*for a very long time humans have used images as proof for things. [...] having an image that leans toward whatever lie someone might be telling is a way to like add credit where it doesn't exist.*" - FG4_P14

Some participants also mentioned the potential of AIGIs to be used by criminals, with examples being blackmailing, creating pornographic material, or scams, e.g., by selling fake products.

Other concerns were related to creative professions, in the sense that people might pretend to be skilled artists by not disclosing the use of AI. Moreover, participants identified copyright issues if GenAI copies or "steals" the style of other artists without permission. Participants also criticized that models might be trained on biased data, which they will later reproduce and cause AIGIs, e.g., to communicate unrealistic (beauty) standards. Another concern was search results or feeds being flooded with AIGIs.

We additionally asked about strategies and experiences with spotting AIGIs. Participants overwhelmingly recognized generated images via visual cues. Two prominent examples were mistakes in the images, e.g., humans with missing or unnatural body parts, and the softness or glossiness of AIGIs. Some participants also talked about context cues, e.g., content was regarded as absurd or too unrealistic to happen in real life. While we had a few participants that stated to find recognition either easy or hard, other participants were more undecided in their judgment. For them, the quality and content of an image influenced how well it can be spotted. A recurring theme was that recognition will become increasingly difficult if GenAI advances further over time. Participants were especially concerned about the vulnerability of loved ones with lower media literacy, e.g., elderly people:

> "*[...] and I'm afraid that someday she [my grandmother] will be misleading by some sort of page [...] and [...] it will end up very bad.*" - FG3_P8

4.2 The Good: AI Labels Can Help to Avoid Deception

When asked about their experiences with AI labels, most participants reported that such labels had not played a significant role in their online activity so far. Some had already experienced AI labels on social media, and a few participants were already confronted with the labeling system when uploading their content. The rest of our participants had not yet encountered AI labels for AIGIs, some had never even heard of them.

While they had little experience so far, our participants generally liked the concept of AI labels. Most of our participants found labels to be helpful, as one participant put it:

> "*I think they are great. So you don't have to question yourself whether something is real or not. Especially if you're not like very tech-savvy.*" - FG1_P4

Adding to that, some participants concretely stated that labels would also help users in identifying misinformation. As AI advances, a few participants found that labels will be a necessity in the future to still be able to distinguish AIGIs from human-made images. Seeing great value in them, some participants found that labeling AIGIs should be mandatory. Participants also stated that, as a side effect, labels could increase the general awareness of AI. They considered this especially valuable for people who are less familiar with the topic of AI, such as elderly people. It was hypothesized that the effect could even persist if not all AIGIs are correctly labeled. We also found quite positive sentiments towards labels in our survey. At its end, we asked all participants in our treatment groups whether they would like to see those labels on real-world social media platforms. An overwhelming majority of 75.7% (679) answered positively, 10.8% (97) were undecided, and only 13.5% (121) negative, showing the great potential of AI labels.

To probe previous knowledge and comprehension of AI labeling, we asked the focus group participants what labeling mechanisms they were aware of or could imagine. Many participants talked about detection, often assuming that it would involve AI. Some participants were aware of visible and invisible watermarks, which was sometimes associated with metadata: "*[...] a way to put that metadata in the actual image itself, but would be invisible to humans*" (FG4_P15). Interestingly, the straightforward approach of embedding metadata into the file was only mentioned by few participants. Others considered the option to disclose the use of AI when uploading media to a social network or to implement a mechanism similar to X's *Community Notes* to flag AIGIs.

Critically, half of our focus group participants stated that if an image is labeled as AI-generated, they would perceive the content more negatively, which reflects the ambivalent attitude of our participants towards AIGIs. A few participants said they would not engage with content if they knew it was made using GenAI. Others stated that disclosing AI usage could make a site or company appear less credible. Besides the general rejection of AI content, disapproval could also be content-sensitive. News, ads, or art were examples where AI labels could lead to a negative perception. On the other hand, a few participants mentioned that labels would not change their perception of content that they consider benign. As one participant noted: "*something creative or something that would, yeah, I would not see in real life. That's where I like to see an AI image*" (FG1_P4).

4.3 The Bad: AI Labels Are Full Of Pitfalls

While seeing their value, most participants did not consider labels a perfect solution and discussed their limits. We report the main issues our participants identified. While some were overarching concerns about AI labeling, others referred to the currently most relevant labeling mechanisms (self-disclosure, detection, and metadata).

Standardization. Many participants were concerned about standardization, finding it hard to decide which AIGIs should or should not be labeled and what labels should look like. A few participants discussed whether content should be labeled depending on the context, specifying that disturbing or contentious images need a label. But they found it hard to decide which content would fit into this category and were concerned about hidden edge cases:

> "*What if you generated an image of your cat jumping out the roof and a child sees it and wants to reproduce it? There should be also here a responsibility to use a label.*" - FG2_P6

Therefore, most of our participants decided that all AIGIs need a label. When discussing labeling rules for images that are only partly generated or edited using AI, participants found the decision to be even harder. A few participants found labels unnecessary if AI was only used for minor edits, like background enhancements or filters, as "conventional" image filters have been used for years without a label. On the other hand, some participants found that all edits involving AI need a label, as it is hard to quantify if the meaning of an image has changed, even if the manipulation only affects a few pixels. To account for differences between entirely generated images and those only edited using AI, some participants proposed to use different labels to distinguish them. However, one participant wondered if this would make the labeling system too complex.

Looking at the broader picture, some participants stressed the importance of consistency regarding similar rules across different social media platforms and countries. Even considering only a single platform, they anticipated negative implications if only some AI images were labeled while others were not. Overall, participants expressed a clear preference for simple and comprehensible rules for labeling.

Abuse of Power. Half of our participants were concerned about the power a platform or authority would have if in charge of regulating and enforcing a labeling system. It was suspected that companies might try to push their agenda:

> "*I don't know if I necessarily trust a platform to do the right thing because I've heard of many instances where they're like, oh we're gonna try to do the right thing and [...] they don't.*" - FG2_P6

Participants were also concerned about platforms using detectors to label AIGIs, which were perceived as particularly non-transparent. Participants had general doubts about their reliability, partly due to experiences with other GenAI tools: "*But AI assessing AI, it's probably not reliable. [...] like ChatGPT is not always reliable.*" (FG1_P2). A recurring theme was that the performance of such a detector would depend on the data it is trained on, allowing the responsible party to influence what is labeled and what is not. This uncertainty regarding how the AI will behave or evolve even caused discomfort among a few participants: "*I don't like thinking about this, this is scary.*" (FG4_P15). Due to the lacking trust in authorities, participants' opinions on who they would like to create and enforce the labeling rules varied greatly. The answers included the government, social media platforms, community efforts, central organizations, providers of AI services, and content creators themselves.

Dishonest Users. Most of our participants questioned other users' honesty and were uncomfortable with mechanisms that solely relied on it. This issue was especially discussed for self-disclosure since users could easily lie about using AI. As one participant put it: "*self-disclosure is like probably the least trustworthy because [...] it would be almost impossible to tell if someone is being honest*" (FG5_P16). But participants were also concerned about the intentional removal of metadata, with some of our participants being genuinely surprised that this can be easily achieved by, for instance, taking a screenshot. Other participants assumed that criminals could simply use (custom) GenAI tools that do not insert metadata, thus bypassing detection.

Usability. Some of our participants identified usability issues with the suggested labeling mechanisms. A recurring theme was that AIGC might be accidentally not labeled, e.g., if a platform relies on self-disclosure, a user might forget to add a label. A few participants even encountered difficulties disclosing the use of AI themselves, suggesting that users are not well informed about the labeling mechanism and its effect:

> "*For Instagram [...] there is an option to say it's [...] AI content, which I have tried but I don't know how to operate it maybe. And I'm like, okay there is no big difference whatever I try to do with that option.*" - FG1_P3

However, usability issues can also affect metadata, as it can be unintentionally removed, e.g., when uploading images to platforms not supporting such metadata. Another concern, related to AI detectors, was that, if the model outputs a probability of an image being AI-generated, this might be difficult to interpret. However, one focus group found that including such a percentage in labels would make them more informative, allowing the user to interpret them.

Mislabeling. Some participants were concerned about mislabeling and its consequences, including false positives and negatives or images that were not classified. In this respect, a few participants worried that users would over-rely on the labels:

> "*what about all the AIGIs that aren't classified? [...] I'm worried that it will just lead to more misinformation if people just blindly trust that anything that's not tagged as AI-generated is actually, you know, valid or real.*" - FG2_P5

One focus group highlighted the necessity of an appealing system, as labels could be assigned incorrectly. We discuss mislabeling and its consequences in more depth in Section 4.4.

4.4 The Ugly: Mislabeling Might Erode Trust in Labels

Most participants found mislabeling to be problematic. However, the level of concern differed between the two types of possible labeling errors. Half of the participants found unlabeled AIGIs (false negatives) to be more concerning than wrongly labeled human-made images (false positives), mainly due to their potential to misinform and cause confusion or fear. Some participants noted that the consequences strongly depend on the image itself and the context it is shared in, with common examples being AIGIs involving politicians or celebrities.

In contrast, the implications of mislabeled human-made images were considered not as severe. A few participants reasoned that uncovering this kind of mislabeling is easier due to common knowledge or the existence of other images of, e.g., the same event, making them "*easier to authenticate*" (FG4_P14). Nevertheless, one group agreed that mislabeled historic photos could make people question past events, like 9/11 or the Holocaust:

> "*That's another doom I haven't thought about until now. Unravelling the implication on history books or politics. Well, that's [a] huge mess [...] we are getting in.*" - FG2_P6

Another concern was the potential reputational damage, e.g., for artists being falsely accused of not creating original work or for politicians being perceived as dishonest. Noteworthy, a few participants were even surprised by the possibility of wrongly labeled human-made images, suggesting that the dangers of those might be less present or tangible.

Some participants considered both mislabeling cases to be equally dangerous, with one participant stating "*It's just as bad. We have to be able to tell what's reality and what's not and it's just as bad to me.*" (FG3_P11). With regard to other images of an event, one participant emphasized that the implications of mislabeled human-made images would be equally harmful if only a single image supported the claim. We also asked our survey participants which type of mislabeling they considered worse. The majority found both equally bad (426, 31.5%), followed by false negatives (301, 22.2%) and false positives (156, 11.5%). Only 1% (14) of our participants considered mislabeling not to be a problem.

We finally asked participants in our focus groups how mislabeling would affect their trust in the labeling system. While many participants stated that observing mislabeled images would make them lose confidence in the label, they had different views on what degree of mislabeling is acceptable. While some could tolerate the occasional mislabeling of images, other participants would not: "*To me, if it even happens once then no trust*" (FG3_P10). Intrigued by this ambiguity, we added a question on how mislabeling affects trust to our survey questionnaire. The results in Figure 2 show that for most participants, even occasional mislabeling of AI-generated or human-made images would lead to a loss of trust.

Interestingly, a few participants stated they would lose trust more quickly if "obvious" AIGIs are mislabeled, hinting towards a misconception regarding the functioning of labeling mechanisms: "*[if] you can clearly tell something is like created by AI and it's not labeled [...] it's kind of like, okay, is this really working?*" (FG3_P9).

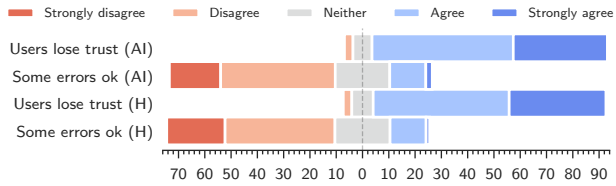


Figure 2: Participants' opinion on how mislabeled AI-generated (AI) and human-made (H) images affect users' trust. Claims were "Users lose trust in the labeling system if they become aware of such mislabeling." and "It is not a problem if such mislabeling only happens once in a while."

Beyond the labels themselves, a few participants found that mislabeling could damage the trust towards the post's source, e.g., a newspaper posting an image alongside a headline: "*Errors do happen, but if they happen multiple times you start questioning about it.*" (FG2_P7). Especially if otherwise credible institutions shared mislabeled AIGIs, this would strongly erode the participants' trust in them.

5 Effects and Side Effects of AI Labels

In this section, we present the results of our survey. We initially describe the hypotheses stated in our pre-registration. We then report our findings on the effects of AI labeling on users' judgments.

5.1 Hypotheses

In our pre-registration, we hypothesized that AI labels might affect participants' judgments in two ways, which we summarize in the following. Note that hypotheses named *HA* compare the control group with both treatment groups (Contrast 1), while hypotheses named *HB* explore the differences between the labeling and mislabeling group (Contrast 2).

From a security perspective, the ideal outcome would be if labels nudged participants towards a state of mind that focuses more on the veracity of a post's claim, regardless of whether it is labeled. Suppose such an increased *truth focus* occurs. In that case, we assume that the treatment groups' sensitivity and accuracy in judging the truthfulness of the information will be higher than those of the control group (*HA1*). It is also possible that this nudging only affects labeled posts, in which case the truth focus and, thus, sensitivity/accuracy would be higher for such posts only (HA2). Moreover, if mislabeling affects participants' judgment, this would cause sensitivity and accuracy to differ compared to the labeling group (HB1). This leads us to three hypotheses for the truth focus. For every hypothesis, accuracy is analyzed concerning the veracity of images.

HA1: Sensitivity and accuracy will be higher in the treatment groups compared to the control group.

HA2: Sensitivity and accuracy will be higher for labeled posts in the treatment groups compared to the control group.

HB1: Sensitivity and accuracy will differ as a function of Image Type between labeling and mislabeling groups.

Alternatively—and less desirable regarding security—participants in the treatment groups might judge claims based on the label.

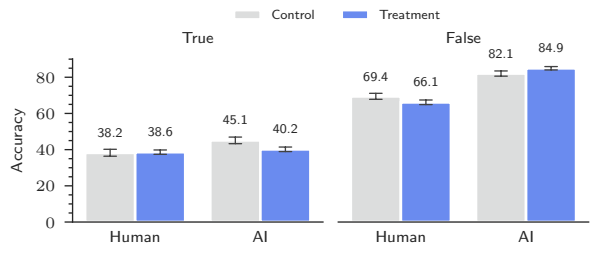


Figure 3: Mean accuracy with 95% confidence intervals, separated by Image Type (human vs. AI), Veracity (true vs. false), and Group (control vs. treatment).

For example, posts with labeled images are rated as false, while those with unlabeled images are rated as true, regardless of their veracity. If mislabeling affects posts in the AI/true and AI/false subsets similarly, the hit and false alarm rates should increase by the same amount. However, participants in the treatment groups should have higher accuracy for AI/false posts (and lower for AI/true). In contrast, since participants might consider the absence of a label as a sign that a claim is true, we would observe the opposite effect for posts in the human/true and human/false subsets (HA3). Comparing the labeling and mislabeling groups should reveal a similar pattern, with no difference in sensitivity but an effect on accuracy (HB2). In summary, we have two hypotheses in case participants tend to judge posts based on their labels. Again, accuracy is analyzed concerning the veracity of images.

HA3: Accuracy will be higher for AI/false and human/true posts but lower for AI/true and human/false posts in the treatment groups compared to the control group.

HB2: While sensitivity will show no difference, the accuracy will differ between the labeling and mislabeling group as a function of Image Type.

We analyze our hypotheses in the following sections. Following our pre-registration, we also investigated the influence of AI labels on confidence and response bias [57] but did not find a statistically significant effect for our pre-registered analysis. We provide the analysis results in Appendix C.1.

5.2 Participants Do Not Generally Make Better Judgments in the Presence of Labels

First, we analyze the effects of labeling on sensitivity. To investigate HA1, HA2, and HB1, we run an ANOVA consisting of the factors Group (C, L, M) × Image Type (AI, human) with sensitivity as the dependent variable and planned Helmert contrasts for the main effect of Group and the Image Type × Group interaction. The analyses reveal an effect for the first Helmert contrast, $t(1351) = -2.00$, $p = .046$, $d = .05$, indicating that sensitivity is *higher* in the control group ($M = .47$, $SD = .76$, 95% $CI = (.42, .52)$) as compared to both treatment groups ($M = .41$, $SD = .75$, 95% $CI = (.38, .45)$). Contrary to the prediction in HA1, labeling made it overall *harder* (instead of easier) to distinguish between true and false claims. Looking at the participants' hit and false alarm rates, we find that the lower sensitivity is driven by a lower hit rate in the treatment groups compared to the control group (.41 vs. .43) with no difference in false alarms (both .28). The effects for sensitivity predicted in the other hypotheses (HA2 and HB1) do not reach statistical significance ($p > .57$).

Next, we study the effect of labels on accuracy through GLMM analyses (see Table 3 for model parameters). To investigate whether an increased truth focus caused by the presence of labels increases participants' accuracy (HA1), we compare a model (Model HA1) with the first Helmert interaction contrast of Group × Veracity (including all main effects) with the simpler model (Model HA1comp) only excluding the two-way interaction. Although the model comparison ($AIC_{\text{Model HA1}} = 37\,027.82$, $AIC_{\text{Model HA1comp}} = 37\,028.56$) suggests a better fit for the model of interest, the first Helmert interaction contrast is not significant ($p = .10$) and thus does not support the hypothesis that labeling leads to an increased truth focus.

In summary, our analysis regarding sensitivity and accuracy speaks against the general truth focus hypothesis. Instead, sensitivity is, contrary to the prediction, even worse for the treatment groups than for the control group. Inspection of the hit and false alarm rates shows that participants, in the context of labeling, rated fewer true posts as true, without a generally more conservative response pattern (i.e., being more hesitant to rate a claim as true). This effect is independent of the image type in the post.

5.3 Participants Tend to Judge a Post by Its Label

To investigate the response behavior of our participants regarding HA3, we compare a model (Model HA3) with a three-way interaction of the first Helmert contrast of Group × Image Type × Veracity (including all main effects) with the simpler model (Model HA3comp) only excluding the three-way interaction. Model HA3 explains the data better than the comparison model ($AIC_{\text{Model HA3}} = 36\,997.86$, $AIC_{\text{Model HA3comp}} = 37\,027.70$). This fits our predictions for HA3, according to which participants' accuracy is affected differently depending on the combination of Image Type and Veracity (see Figure 3).

Following our pre-registration, we calculate simple comparisons for each Image Type × Veracity combination to better understand the significant three-way interaction effects. For posts containing AIGIs, true posts show lower accuracy in the treatment groups compared to the control group, $z = 4.20$, $p < .001$, while the accuracy is higher for false posts, $z = -3.35$, $p < .001$. For human-made images, the accuracy for posts containing false claims is lower in the treatment groups compared to the control group, $z = 3.11$, $p = .002$. Human/true posts do not differ between control and treatment groups, $z = -0.34$, $p = .736$. Accordingly, we do not find evidence for a limited truth focus that only applies to labeled posts (HA2).

The detailed results for participants' accuracy indicate that people (over-)rely on the AI labels when judging claims as true or false, supporting HA3. Posts containing AIGIs are often perceived as false, regardless of the veracity. The effect becomes apparent when comparing the mean accuracy between control and treatment groups, as shown in Figure 3. An accuracy increase for AI/false posts from 82.1% to 84.9% can be interpreted as follows: In the control group,

Table 3: GLMMs to investigate hypotheses on accuracy. The models' suffixes indicate the tested hypotheses.

	Model HA1				Model HA3				Model HB2			
Predictors	Odds Ratios	CI (95%)	z-value	*p*	Odds Ratios	CI (95%)	z-value	*p*	Odds Ratios	CI (95%)	z-value	*p*
(Intercept)	3.67	2.43–5.54	6.18	**<0.001**	3.68	2.56–5.30	7.01	**<0.001**	3.68	2.55–5.30	6.99	**<0.001**
Veracity	0.17	0.10–0.31	-5.89	**<0.001**	0.17	0.10–0.29	-6.68	**<0.001**	0.17	0.10–0.29	-6.66	**<0.001**
Contrast 1	1.00	0.97–1.02	-0.30	0.761	1.01	0.98–1.04	0.52	0.602	0.98	0.96–1.00	-2.16	**0.031**
Contrast 2	0.99	0.96–1.03	-0.45	0.649	0.99	0.96–1.03	-0.46	0.649	1.06	1.01–1.12	2.41	**0.016**
Veracity × Contrast 1	0.97	0.94–1.01	-1.66	0.097	0.96	0.93–1.00	-2.18	**0.029**				
Image Type					0.63	0.43–0.90	-2.53	**0.011**	0.63	0.43–0.90	-2.52	**0.012**
Veracity × Image Type					1.44	0.86–2.40	1.39	0.166	1.44	0.86–2.41	1.38	0.168
Image Type × Contrast 1					0.94	0.91–0.96	-4.68	**<0.001**				
Veracity × Image Type × Contrast 1					1.11	1.07–1.15	5.66	**<0.001**				
Veracity × Contrast 2									0.90	0.85–0.96	-3.18	0.001
Image Type × Contrast 2									0.91	0.86–0.95	-4.02	**<0.001**
Veracity × Image Type × Contrast 2									1.20	1.12–1.27	5.62	**<0.001**

Coding of predictors: Veracity (true = 1, false = -1), Image Type (human = 1, AI = -1), Contrast 1 (control = -2, treatment = 1), Contrast 2 (labeling = 1, mislabeling = -1)

out of 1 000 participants, 821 would recognize the claim of the post as false, while it would be 849 people in the treatment groups. Thus, labeling would result in 28 more participants correctly rating the post's claim as false. Contrary to our assumptions, the judgment of human/true posts was not affected significantly. However, the decreased accuracy for human/false posts in the treatment groups might suggest an implied truth effect [51], according to which participants interpret the absence of a label as an indication that a claim is true.

To explore an alternative explanation of our findings, we conducted a post hoc analysis of the response bias [57], i.e., the tendency to judge a claim as true or false regardless of its veracity, depending on the image type. We found that participants in the treatment groups hesitated more than participants in the control group to classify a claim as true if it was accompanied by an AIGI. However, if a claim was presented alongside a human-made image, participants in the control groups were more hesitant to rate it as true than participants in the treatment groups. These results could indicate that participants do not simply judge a claim as false because it is labeled, but that labels lead to a more conservative response behavior. On the other hand, participants might be more willing to judge an unlabeled post as true if they have seen other labeled posts. The details of our analysis can be found in Appendix C.1.

5.4 Mislabeled AI-Generated Images Are More Often Judged as True

To investigate whether participants over-rely on labels in the face of mislabeling (HB2), we compare a model (Model HB2) with a three-way interaction of the second Helmert contrast of Group × Image Type × Veracity (including all main effects) with the simpler model (Model HB2comp), which only excludes the three-way interaction. Our results indicate that the three-way interaction model explains the data better than the comparison model, indicating that this interaction is important in explaining our data ($AIC_{\text{Model HB2}} = 36\,993.40$, $AIC_{\text{Model HB2comp}} = 37\,022.88$). To better understand the significant three-way interaction, we run simple comparisons. AI/true posts show lower accuracy in the labeling group compared to the mislabeling group ($z = 4.18,\ p < .001$). For AI/false posts the accuracy is higher in the labeling group compared to the mislabeling group,$z = -4.28,\ p < .001$. For human/true posts ($z = -1.34,\ p = .18$) and human/false posts ($z = 1.13,\ p = .26$), accuracy does not differ between the labeling and the mislabeling group.

Overall, we find evidence for our hypothesis HB2 and consequently reject HB1. People still rely on labels in the case of mislabeling. Participants associate the absence of a label (if AIGIs are not labeled) as an indication that the claim is true, so they judge them less often as false. However, according to our hypothesis, we expected the comparisons for true and false human posts between the labeling and mislabeling groups to become significant as well. We assume that mislabeling might have a more substantial impact on AI images than on human images.

6 Discussion

In this section, we discuss our qualitative and quantitative findings in the context of our three research question. First, we derive recommendations for the successful deployment of labels based on users' opinions, expectations, and concerns about AI labeling. Second, we discuss to what extent AI labels can protect users against misinformation, considering their side effects. Lastly, we answer how the consequences of mislabeling impact the practical use of AI labels on social media platforms.

6.1 RQ1: What Are Users' Opinions, Expectations, and Concerns About AI Labeling?

Participants were overwhelmingly in favor of AI labels. The plethora of threats that AIGIs entail, coupled with the continuously improving visual quality, made participants aware that relying solely on visual recognition might not be sufficient to detect AIGC. Therefore, they considered AI labels a suitable tool to spot AIGIs easily. This puts AI labels in a favorable situation: Users see their merits and are willing to adopt them, which is not the case for other security tools, e.g., password managers [63]–[65], that do not have an obvious everyday value. While AI labels are primarily intended as a transparency mechanism, they were also perceived as a valuable safeguard against misinformation.

However, AI labeling needs to be thoughtfully designed and implemented. Without even touching on the visual design, we have

already identified several underlying concerns that have the potential to steer users away from labels: Participants questioned the rules about what is labeled and what is not, and were especially suspicious of platforms abusing their power, pushing certain narratives by selectively enforcing labels. Moreover, they were concerned about mislabeling and questioned the purpose of mechanisms that relied on others' honesty. Lastly, we identified usability issues that must be addressed for a successful adoption. As we found users' trust to be a valuable but fragile resource, we recommend considering the following points for a successful deployment of AI labels:

Simplicity and Consistency. Participants want simple and comprehensible labeling policies. As such, we found evidence that all AIGIs should be labeled. This contradicts existing legislation (see Section 2.1), which often only demands labels for contentious content. Users might expect a different approach to labeling partly generated or edited images. However, more research is needed to investigate this further. Participants stressed that labeling rules need to be consistent, not only within a platform but also between platforms. However, the current landscape of labeling policies varies drastically (see Section 2.2). While consensus between platforms might be hard to reach, platforms should at least deploy consistent rules on their platforms, e.g., not allowing labeling exceptions or using different labels for different kinds of AIGIs.

Transparency and Correctness. A central concern of our participants was a potential abuse of power. If users have reasons to question the neutrality of the labeling system, they will likely distrust it. To prevent this, platforms need to transparently inform about their labeling rules and which mechanisms are used. This is especially relevant for detectors, as users might not comprehend their decisions, which could raise concerns about intentional biases from training data. Moreover, participants were especially critical of self-disclosure, knowing malicious users would not disclose their AI usage. Thus, platforms should disclose the capabilities and weaknesses of their labeling system. Since mislabeling can quickly erode users' trust, occurrences of mislabeling should be transparently addressed.

Usability. When AI labels are introduced to a platform, they must also be introduced to its users. Many participants did not notice AI labeling, even on platforms that had already deployed it. The platform-specific labeling procedures must be made clear so that users know what is asked of them when uploading their content.

6.2 RQ2: How Does AI Labeling Affect Users' Perception of True and False Claims With and Without AI-Generated Images?

A central argument for AI labeling is to protect the public against AI-generated misinformation. Our focus group discussions confirmed that users know the diverse harms of AIGC and would appreciate labels to mitigate the associated risks. However, our survey showed that AI labels can only partially meet the expectations placed on them. While we could confirm previous work [21] that found labels reduce the belief in misleading claims supported by AIGIs, our study design revealed statistically significant side effects.

Most critically, the presence of labels made participants more susceptible to misleading posts featuring human-made images that were, e.g., maliciously taken out of context or cropped. Despite the growing concerns about the malicious use of GenAI, this "conventional" misinformation is still a pressing problem [19]. Our results could be explained through an implied truth effect [51], suggesting that, in the presence of labels, users are more likely to believe the associated claim is valid if a post is not labeled. However, in this case, users would mistake AI labels for misinformation warnings, an observation already discovered in the context of provenance indicators [66]. Another explanation could be that AI labels distract users, making them look out for the novel dangers of AIGC while forgetting about "conventional" misinformation. If this is the case, the security risks of AI labels might be higher than their (current) reward. We therefore call upon future work to further investigate *if* AI labels influence the perception of such misinformation in real-world scenarios and *why* AI labels cause this change in perception.

Our focus group findings, as well as related work [21], [67], [68], indicate that users start to question AIGC if it is disclosed as such, which could explain our second side effect: In the presence of labeling, participants' belief in true claims accompanied by AIGIs was reduced. As long as users are skeptical towards AIGC, transparency of AIGC might lead to aversion to the underlying content. Again, AI labels might be perceived as misinformation warnings for these users. Currently, most news agencies forbid the use of GenAI to illustrate news articles with realistic-looking AIGIs [69]. However, more abstract AIGIs are increasingly being used as an alternative to stock photos, especially for less controversial articles [56]. Given that the lines regarding the acceptable use of GenAI are getting blurry, we argue that future work on AI labeling should consider legitimate uses of AIGC. While our study found that AI labels are perceived as a crucial tool, we argue that AI labels can have considerable adverse side effects on non-malicious AIGC as long as the usage of GenAI remains controversial.

Regardless of the potential side effects, we found that AI labels' constructive effect was moderate overall, similar to related work [67]. Taken together, we argue that such labels can only be one cornerstone in the fight against AI-generated misinformation and should be treated accordingly by legislators and social media platforms. In this light, Meta's decision to label instead of remove potentially misleading AIGC [70] should be critically reviewed. Our findings suggest that labeling cannot fully eliminate the risks of deceptive AIGC, so removing harmful content should be prioritized.

6.3 RQ3: How Does Mislabeling Interfere With the Efficacy of and Trust Towards AI Labeling?

We found mislabeling to be a significant threat to the success of AI labeling. Our experiment showed that users rely on labels when judging a post's claim, regardless of whether an image is correctly labeled or mislabeled. Thus, in a context where labels are present, they are more likely to fall for misinformation conveyed through an unlabeled AIGI. However, our focus groups revealed that mislabeling can have even more far-reaching implications. Our results strongly suggest that users lose trust in the labels if they encounter mislabeling. We argue there is little room for errors, as this loss can

happen after seeing a single mislabeled image. Currently, participants are open to labels, but once their trust is gone, it might be much more difficult to restore it and convince users of the benefits of AI labels.

Moreover, platforms must consider the social risks of false positives if content creators are confronted with (supposedly) unjust accusations. Meta's original label design, which stated that content was "Made with AI", was changed to "AI Info" after backlash from users who considered their images to be falsely flagged [71]. While we argue that users might still not be content with their images receiving this label, it is also unclear how much this label still informs about AI usage. Instead, if mislabeling cannot be avoided, it needs to be openly and transparently processed, including an appeal system, to give users the certainty that they can generally trust the labeling system.

7 Related Work

This section presents existing work on the labeling of AIGC. We also put previous work in relation to our study. Since AI labels often aim to inform users about deceptive media, we initially discuss previous findings related to misinformation warnings.

7.1 Misinformation Warnings

In their review of previous research, Martel *et al.* [72] found that misinformation warnings, presented alongside the misinformation, can be used as an effective tool to combat deceptive media. Investigating the effect of warning labels on Twitter, Papakyriakopoulos *et al.* [73] found that, overall, labels did not impact the interaction with posts but that contextual or well-explained warnings could reduce it. However, misinformation warnings are not without side effects. Hoes *et al.* [74] investigated the effectiveness of three misinformation intervention strategies. According to their findings, the strategies reduced participants' belief in misinformation, but they also made participants more suspicious of authentic information. Adding to that, Pennycook *et al.* [51] found that misinformation labels can lead to an *implied truth effect*, meaning that, in the presence of labels, users trust content that is not labeled more, as they assume that it passed a fact check. Uncovering additional concerns, Hameleers *et al.* [75] studied the consequences if misinformation labels are maliciously assigned and found that they can reduce the credibility of authentic content.

This literature uncovering side effects of misinformation warnings led us to explore the side effects of AI labels and how users would engage with them. To account for potential effects, we chose a design that includes human-made and AI-generated images, as well as true and false claims.

To increase the effectiveness of misinformation warnings, previous work investigated multiple design decisions. While Kaiser *et al.* [76] compared contextual and interstitial warnings against disinformation and found interstitial warnings to be more effective, Sharevski *et al.* [77] investigated contextual and iconography designs. While a deterrent effect of AI labels must be carefully weighted, as benign content is also labeled, we assume that different design decisions will also have a huge impact. We call on future work to investigate those effects, as, in our work, we deem it important to first investigate fundamental issues and user perceptions of AI labels.

7.2 Labels for AI-Generated Content

To lay a first foundation, Epstein *et al.* [50] investigated the understanding of textual labels for AIGC. While terms such as "AI Generated", "Generated with an AI Tool", or "AI Manipulated" were correctly associated with media created using generative AI, participants considered "Deepfake" or "Manipulated" content to be intentionally misleading. Altay *et al.* [67] investigated AI labels for news headlines and accounted for human-made and AI-generated content, as well as for true information and misinformation. They found that trust in labeled headlines is reduced, even if they are true or authentic, as participants believed that the whole text was written by AI. While they used a similar design to ours, their research focused on AI-generated news headlines.

While Wittenberg *et al.* [21] investigated the effect of AI labels for AIGIs, their stimuli set contained only images that were AI-generated *and* misleading. These images were embedded into a simulated social media post, together with different variants of labels. The authors found that the presence of labels generally reduced belief in AIGIs. However, by design, the experiment could not investigate effects of labels on benign posts with AIGIs or misinformation accompanied by human-made images. We consider this type of posts important, as previous research illustrates that labels do not only have an effect on misinformation but also on benign AIGC. Toff *et al.* [68] evaluated the effect of labeling regarding the trust in journalistic content. Participants considered a news article to be less trustworthy if it is labeled as AI-generated. Lim *et al.* [78] investigated how participants reacted to AI-generated health prevention messages if labeled as such. They found that disclosing the source had a negative impact on participants' assessment by a small but significant amount. Additionally, two similar studies [79], [80] on the detectability of deepfake videos showed that warnings did increase skepticism towards shown videos. However, due to the inability of humans to reliably discern fake videos from authentic ones, this effect existed regardless of whether the video is a deepfake. Concerning this inability, Rae [81] studied whether labels for text matter in a future where content created by AI cannot be distinguished from that written by a human. They found that participants had more negative feelings towards content creators when they believed that AI was involved and were less satisfied.

Our research confirms previous work that AI labels can reduce participants' belief in misinformation, but also that participants assess benign content more negatively if it is labeled. Critically, differing from Altay *et al.* [67], we found that, in the presence of labels, participants were more susceptible to misleading posts containing unlabeled human-made images. This could be an indication of an implied truth effect [51].

Moreover, we qualitatively assessed our participants' perception and acceptance of labels and uncovered problems that could impair their trust. While Ali *et al.* [82] developed 149 questions regarding transparent AI disclosure within a participatory workshop, to the best of our knowledge, we are the first to set out to uncover

user opinions about AI labels, including key issues that need to be addressed to adopt labels for AIGIs successfully.

8 Conclusion

In this work, we study the implications of labels for AIGIs. We conducted five focus groups and a pre-registered online survey with over 1 300 U.S. and EU participants to investigate users' perceptions and measurable effects on reducing the risks of AI-generated misinformation. We found that users overwhelmingly favor AI labels as a transparency mechanism *and* a promising tool to recognize misinformation. However, we also uncovered several underlying issues that need to be addressed, including mislabeling, which could erode users' trust.

Our survey measured the effect of AI labels on human-made and AI-generated images conveying both true and false information. While labels can decrease users' belief in misinformation if paired with AIGIs, we also uncovered that true information illustrated with a labeled AIGI was more frequently dismissed as false. Moreover, labels made users perceive "conventional" misinformation (without any involvement of GenAI) as more credible. Our findings suggest that AI labels can be a simple and effective tool for creating transparency and mitigating the risks of AI-generated misinformation. However, they must be carefully implemented to avoid undesired side effects.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2092 CASA - 390781972. Moreover, this work was supported by the German Federal Ministry of Research, Technology and Space under the grants AIgenCY (16KIS2012) and UbiTrans (16KIS1900). This research was partially funded by VolkswagenStiftung Niedersächsisches Vorab – ZN3695.

References

[1] E. Griffith and C. Metz, "A new area of A.I. booms, even amid the tech gloom," *The New York Times*, 2023, https://www.nytimes.com/2023/01/07/technology/generative-ai-chatgpt-investments.html.

[2] J. Damiani, "A voice deepfake was used to scam a CEO out of $243,000," *Forbes*, 2019, https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/.

[3] D. Lepido and Bloomberg, "Ferrari exec foils deepfake attempt by asking the scammer a question only CEO Benedetto Vigna could answer," *Fortune*, 2024, https://fortune.com/2024/07/27/ferrari-deepfake-attempt-scammer-security-question-ceo-benedetto-vigna-cybersecurity-ai/.

[4] H. Chen and K. Magramo, "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'," *CNN*, 2024, https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.

[5] K. Conger and J. Yoon, "Explicit deepfake images of Taylor Swift elude safeguards and swamp social media," *The New York Times*, 2024, https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html.

[6] Federal Bureau of Investigation (FBI), *Malicious actors manipulating photos and videos to create explicit content and sextortion schemes*, https://www.ic3.gov/PSA/2023/PSA230605, 2023.

[7] H. Jingnan, "AI-generated images have become a new form of propaganda this election season," *NPR*, 2024, https://www.npr.org/2024/10/18/nx-s1-5153741/ai-images-hurricanes-disasters-propaganda.

[8] T. Hsu and S. L. Myers, "A.I.'s use in elections sets off a scramble for guardrails," *The New York Times*, 2023, https://www.nytimes.com/2023/06/25/technology/ai-elections-disinformation-guardrails.html.

[9] Center for Countering Digital Hate, *Fake Image Factories: How AI image generators threaten election integrity and democracy*, https://counterhate.com/research/fake-image-factories/, 2024.

[10] N. Ibrahim, "Taylor Swift endorsed Trump in these viral images?" *Snopes*, 2024, https://www.snopes.com//fact-check/taylor-swift-endorsed-trump/.

[11] A. Clayton, "Fake AI-generated image of explosion near Pentagon spreads on social media," *The Guardian*, 2023, https://www.theguardian.com/technology/2023/may/22/pentagon-ai-generated-image-explosion.

[12] P. Verma and G. De Vynck, "AI is destabilizing 'the concept of truth itself' in 2024 election," *The Washington Post*, 2024, https://www.washingtonpost.com/technology/2024/01/22/ai-deepfake-elections-politicians/.

[13] Pew Research Center, *Social media and news fact sheet*, https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet, 2023.

[14] G. Corsi, B. Marino, and W. Wong, "The spread of synthetic media on X," *Harvard Kennedy School Misinformation Review*, 2024, https://doi.org/10.37016/mr-2020-140.

[15] Federal Register, *Executive order on the safe, secure, and trustworthy development and use of artificial intelligence*, https://www.federalregister.gov/d/2023-24283, 2023.

[16] D. Shepardson, "Trump revokes Biden executive order on addressing AI risks," *Reuters*, 2025, https://www.reuters.com/technology/artificial-intelligence/trump-revokes-biden-executive-order-addressing-ai-risks-2025-01-21/.

[17] Y. L. Dan Xuezi, "China releases new labeling requirements for AI-generated content," *Inside Privacy*, 2025, https://www.insideprivacy.com/international/china/china-releases-new-labeling-requirements-for-ai-generated-content/.

[18] J. Morales, "China will enforce clear flagging of all AI generated content starting from September," *Tom's Hardware*, 2025, https://www.tomshardware.com/tech-industry/artificial-intelligence/china-will-enforce-clear-flagging-of-all-ai-generated-content-starting-from-september.

[19] N. Dufour, A. Pathak, P. Samangouei, *et al.*, "AMMEBA: A large-scale survey and dataset of media-based misinformation in-the-wild," *arXiv Preprint*, 2024, https://arxiv.org/abs/2405.11697.

[20] Y. Wang, F. Tahmasbi, J. Blackburn, *et al.*, "Understanding the use of fauxtography on social media," *International AAAI Conference on Web and Social Media (ICSWM)*, 2021, https://doi.org/10.1609/icwsm.v15i1.18102.

[21] C. Wittenberg, Z. Epstein, G. Peloquin-Skulski, A. J. Berinsky, and D. Rand, "Labeling AI-generated media online," *PsyArXiv Preprint*, 2024, https://doi.org/10.31234/osf.io/b238p.

[22] European Parliament, *P9_TA(2024)0138 Artificial Intelligence Act*, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf, 2024.

[23] T. Gils, "A detailed analysis of article 50 of the EU's Artificial Intelligence Act," *SSRN Preprint*, 2024, https://dx.doi.org/10.2139/ssrn.4865427.

[24] European Parliament, *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act)*, https://eur-lex.europa.eu/eli/reg/2022/2065/oj, 2022.

[25] D. Bartz and K. Hu, "OpenAI, Google, others pledge to watermark AI content for safety, White House says," *Reuters*, 2023, https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/.

[26] T. Hunnicutt, K. Singh, and K. Singh, "Biden says risks posed by AI to security, economy need addressing," *Reuters*, 2023, https://www.reuters.com/technology/biden-says-there-is-need-address-security-economic-risks-posed-by-ai-2023-06-20/.

[27] Sen. Brian Schatz, *AI Labeling Act of 2023*, https://www.congress.gov/bill/118th-congress/senate-bill/2691/text, 2023.

[28] W. Knight, "Under trump, AI scientists are told to remove 'ideological bias' from powerful models," *Wired*, 2025, https://www.wired.com/story/ai-safety-institute-new-directive-america-first/.

[29] Meta, *Misinformation*, https://transparency.meta.com/policies/community-standards/misinformation/, Accessed March 18, 2025.

[30] Coalition for Content Provenance and Authenticity (C2PA), *C2PA technical specification*, https://c2pa.org/specifications/specifications/2.0/specs/C2PA_Specification.html, 2024.

[31] International Press Telecommunications Council (IPTC) Photo Metadata Working Group, *IPTC photo metadata standard 2023.2*, https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata, 2023.

[32] N. Clegg, *Labeling AI-generated images on facebook, instagram and threads*, https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/, Accessed March 19, 2025, 2024.

[33] Meta, *Label AI content on instagram*, https://help.instagram.com/761121959519495/, Accessed March 18, 2025.

[34] X, *X help center: Authenticity*, https://help.x.com/en/rules-and-policies/authenticity, Accessed March 19, 2025.

[35] X, *Notes on media & links*, https://communitynotes.x.com/guide/en/contributing/notes-on-media, Accessed March 20, 2025.

[36] xAI, *Grok image generation release*, https://x.ai/news/grok-image-generation-release, Accessed March 20, 2025.

[37] TikTok, *Community guidelines*, https://www.tiktok.com/community-guidelines, Accessed March 18, 2025.

[38] TikTok, *Partnering with our industry to advance AI transparency and literacy*, https://newsroom.tiktok.com/en-us/partnering-with-our-industry-to-advance-ai-transparency-and-literacy, Accessed March 18, 2025.

[39] YouTube, *Disclosing use of altered or synthetic content*, https://support.google.com/youtube/answer/14328491, Accessed March 19, 2025.

[40] YouTube, *Building trust on YouTube: 'Captured with a camera' disclosure*, https://support.google.com/youtube/answer/15446725, Accessed March 19, 2025.

[41] LinkedIn, *User Agreement*, https://www.linkedin.com/legal/user-agreement, Accessed March 18, 2025.

[42] LinkedIn, *LinkedIn Professional Community Policies*, https://www.linkedin.com/legal/professional-community-policies, Accessed March 18, 2025.

[43] LinkedIn, *Help: False or misleading content*, https://www.linkedin.com/help/linkedin/answer/a1340752, Accessed March 18, 2025.

[44] LinkedIn, *Help: Content credentials*, https://www.linkedin.com/help/linkedin/answer/a6282984, Accessed March 18, 2025.

[45] D. Wessel, C. Attig, and T. Franke, "ATI-S - an ultra-short scale for assessing affinity for technology interaction in user studies," in *Mensch und Computer (MuC)*, https://doi.org/10.1145/3340764.3340766, 2019.

[46] S. Grassini, "Development and validation of the AI attitude scale (AIAS-4): A brief measure of general attitude toward artificial intelligence," *Frontiers in Psychology*, 2023, https://doi.org/10.3389/fpsyg.2023.1191628.

[47] N. McDonald, S. Schoenebeck, and A. Forte, "Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice," *Proceedings of the ACM on Human-Computer Interaction*, 2019, https://doi.org/10.1145/3359174.

[48] J. H. Klemmer, S. A. Horstmann, N. Patnaik, *et al.*, "Using AI assistants in software development: A qualitative study on security practices and concerns," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, https://doi.org/10.1145/3658644.3690283, 2024.

[49] X. Bouwman, H. Griffioen, J. Egbers, C. Doerr, B. Klievink, and M. van Eeten, "A different cup of TI? the added value of commercial threat intelligence," in *USENIX Security Symposium*, https://www.usenix.org/conference/usenixsecurity20/presentation/bouwman, 2020.

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

[51] G. Pennycook, A. Bear, E. T. Collins, and D. G. Rand, "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings," *Management Science*, 2020, https://doi.org/10.1287/mnsc.2019.3478.

[52] K. J. K. Feng, N. Ritchie, P. Blumenthal, A. Parsons, and A. X. Zhang, "Examining the impact of provenance-enabled media on trust and accuracy perceptions," *Proceedings of*

the ACM on Human-Computer Interaction, 2023, https://doi.org/10.1145/3610061.

[53] W. Oremus and P. Verma, "These look like prizewinning photos. They're AI fakes.," *Washington Post*, 2023, https://www.washingtonpost.com/technology/2023/11/23/stock-photos-ai-images-controversy/.

[54] C. Wilson, "Adobe is selling fake AI images of the war in Israel-Gaza," *Crikey*, 2023, https://www.crikey.com.au/2023/11/01/israel-gaza-adobe-artificial-intelligence-images-fake-news/.

[55] A. Nicoud, *Bringing AI to a 400 year old media group*, https://theaudiencers.com/bringing-ai-to-a-400-year-old-media-group/, 2024.

[56] N. Newman, "Journalism, media, and technology trends and predictions 2024," Reuters Institute for the Study of Journalism, Tech. Rep., 2024, https://doi.org/10.60625/risj-0s9w-z770.

[57] C. Batailler, S. M. Brannon, P. E. Teas, and B. Gawronski, "A signal detection approach to understanding the identification of fake news," *Perspectives on Psychological Science*, 2022, https://doi.org/10.1177/1745691620986135.

[58] U. Granziol, M. Rabe, M. Gallucci, A. Spoto, and G. Vidotto, "Not another post hoc paper: A new look at contrast analysis and planned comparisons," *Advances in Methods and Practices in Psychological Science*, 2025, https://doi.org/10.1177/25152459241293110.

[59] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected papers of Hirotugu Akaike*, https://doi.org/10.1007/978-1-4612-1694-0_15, 1998.

[60] J. Tang, E. Birrell, and A. Lerner, "Replication: How well do my results generalize now? The external validity of online privacy and security surveys," in *Symposium on Usable Privacy and Security (SOUPS)*, https://www.usenix.org/conference/soups2022/presentation/tang, 2022.

[61] E. Peer, D. Rothschild, A. Gordon, Z. Evernden, and E. Damer, "Data quality of platforms and panels for online behavioral research," *Behavior Research Methods*, 2022, https://doi.org/10.3758/s13428-021-01694-3.

[62] B. D. Douglas, P. J. Ewell, and M. Brauer, "Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA," *PLoS One*, 2023, https://doi.org/10.1371/journal.pone.0279720.

[63] S. Pearman, S. A. Zhang, L. Bauer, N. Christin, and L. F. Cranor, "Why people (don't) use password managers effectively," in *Symposium on Usable Privacy and Security (SOUPS)*, https://www.usenix.org/conference/soups2019/presentation/pearman, 2019.

[64] H. Ray, F. Wolf, R. Kuber, and A. J. Aviv, "Why older adults (don't) use password managers," in *USENIX Security Symposium*, https://www.usenix.org/conference/usenixsecurity21/presentation/ray, 2021.

[65] S. Amft, S. Höltervennhoff, N. Huaman, Y. Acar, and S. Fahl, "'Would you give the same priority to the bank and a game? I do not!' Exploring credential management strategies and obstacles during password manager setup," in *Symposium on Usable Privacy and Security (SOUPS)*, https://www.usenix.org/conference/soups2023/presentation/amft, 2023.

[66] I. N. Sherman, J. W. Stokes, and E. M. Redmiles, "Designing media provenance indicators to combat fake media," in *International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, https://doi.org/10.1145/3471621.3471860, 2021.

[67] S. Altay and F. Gilardi, "People are skeptical of headlines labeled as ai-generated, even if true or human-made, because they assume full AI automation," *PNAS Nexus*, 2024, https://doi.org/10.1093/pnasnexus/pgae403.

[68] B. Toff and F. M. Simon, ""Or they could just not use it?": The paradox of AI disclosure for audience trust in news," *SocArXiv Preprint*, 2023, https://doi.org/10.31235/osf.io/mdvak.

[69] N. Heer, *So far, A.I.-generated images of current events seem rare in news stories*, https://pxlnv.com/blog/news-publishers-ai-generated-images/, 2023.

[70] Kurt Wagner, "Meta stops removing some AI-generated posts, even if misleading," *Bloomberg*, 2024, https://www.bloomberg.com/news/articles/2024-04-05/meta-to-label-more-ai-generated-posts-instead-of-removing-them.

[71] R. Lawler, "Instagram's 'Made with AI' label swapped out for 'AI info' after photographers' complaints," *The Verge*, 2024, https://www.theverge.com/2024/7/1/24190026/meta-instagram-facebook-made-with-ai-info-label-metadata.

[72] C. Martel and D. G. Rand, "Misinformation warning labels are widely effective: A review of warning effects and their moderating features," *Current Opinion in Psychology*, 2023, https://doi.org/10.1016/j.copsyc.2023.101710.

[73] O. Papakyriakopoulos and E. Goodman, "The impact of Twitter labels on misinformation spread and user engagement: Lessons from Trump's election tweets," in *ACM Web Conference (WWW)*, https://doi.org/10.1145/3485447.3512126, 2022.

[74] E. Hoes, B. Aitken, J. Zhang, T. Gackowski, and M. Wojcieszak, "Prominent misinformation interventions reduce misperceptions but increase scepticism," *Nature Human Behaviour*, 2024, https://doi.org/10.1038/s41562-024-01884-x.

[75] M. Hameleers and F. Marquart, "It's nothing but a deepfake! the effects of misinformation and deepfake labels delegitimizing an authentic political speech," *International Journal of Communication*, 2023, https://ijoc.org/index.php/ijoc/article/view/20777.

[76] B. Kaiser, J. Wei, E. Lucherini, K. Lee, J. N. Matias, and J. Mayer, "Adapting security warnings to counter online disinformation," in *USENIX Security Symposium*, https://www.usenix.org/conference/usenixsecurity21/presentation/kaiser, 2021.

[77] F. Sharevski, A. Devine, P. Jachim, and E. Pieroni, "Meaningful context, a red flag, or both? Preferences for enhanced misinformation warnings among US Twitter users," in *European Symposium on Usable Security (EuroUSEC)*, https://doi.org/10.1145/3549015.3555671, 2022.

[78] S. Lim and R. Schmälzle, "The effect of source disclosure on evaluation of AI-generated messages," *Computers in Human*

Behavior: Artificial Humans (CHBAH), 2024, https://doi.org/10.1016/j.chbah.2024.100058.

[79] J. Ternovski, J. Kalla, and P. M. Aronow, "Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments," *OSF Preprint*, 2021, https://doi.org/10.31219/osf.io/dta97.

[80] A. Lewis, P. Vu, R. M. Duch, and A. Chowdhury, "Deepfake detection with and without content warnings," *Royal Society Open Science*, 2023, https://doi.org/10.1098/rsos.231214.

[81] I. Rae, "The effects of perceived AI use on content perceptions," in *CHI Conference on Human Factors in Computing Systems*, https://doi.org/10.1145/3613904.3642076, 2024.

[82] A. E. Ali, K. P. Venkatraj, S. Morosoli, L. Naudts, N. Helberger, and P. Cesar, "Transparent AI disclosure obligations: Who, what, when, where, why, how," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, https://doi.org/10.1145/3613905.3650750, 2024.

[83] PAI Staff, *Building a glossary for synthetic media transparency methods, part 1: Indirect disclosure*, https://partnershiponai.org/glossary-for-synthetic-media-transparency-methods-part-1-indirect-disclosure/, Accessed August 8, 2024.

[84] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "HiDDeN: Hiding data with deep networks," in *European Conference on Computer Vision (ECCV)*, https://doi.org/10.1007/978-3-030-01267-0_40, 2018.

[85] M. Tancik, B. Mildenhall, and R. Ng, "StegaStamp: Invisible hyperlinks in physical photographs," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, https://doi.org/10.1109/CVPR42600.2020.00219, 2020.

[86] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, https://doi.org/10.1109/ICCV51070.2023.02053, 2023.

[87] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, "Tree-rings watermarks: Invisible fingerprints for diffusion images," *Advances in Neural Information Processing Systems (NeurIPS)*, 2023, https://proceedings.neurips.cc/paper_files/paper/2023/hash/b54d1757c190ba20dbc4f9e4a2f54149-Abstract-Conference.html.

[88] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang, and N. Yu, "Gaussian shading: Provable performance-lossless image watermarking for diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, https://doi.org/10.1109/CVPR52733.2024.01156, 2024.

[89] H. Ci, P. Yang, Y. Song, and M. Z. Shou, "RingID: Rethinking tree-ring watermarking for enhanced multi-key identification," in *European Conference on Computer Vision (ECCV)*, https://doi.org/10.1007/978-3-031-73390-1_20, 2025.

[90] S. Gunn, X. Zhao, and D. Song, "An undetectable watermark for generative image models," in *International Conference on Learning Representations (ICLR)*, https://openreview.net/forum?id=jlhBFm7T2J, 2025.

[91] X. Zhao, K. Zhang, Z. Su, *et al.*, "Invisible image watermarks are provably removable using generative AI," in *Advances in Neural Information Processing Systems (NeurIPS)*, https://proceedings.neurips.cc/paper_files/paper/2024/hash/10272bfd0371ef960ec557ed6c866058-Abstract-Conference.html, 2024.

[92] P. Yang, H. Ci, Y. Song, and M. Z. Shou, "Can simple averaging defeat modern watermarks?" In *Advances in Neural Information Processing Systems (NeurIPS)*, https://proceedings.neurips.cc/paper_files/paper/2024/hash/67b2e2e895380fa6acd537c2894e490e-Abstract-Conference.html, 2024.

[93] H. Zhang, B. L. Edelman, D. Francati, D. Venturi, G. Ateniese, and B. Barak, "Watermarks in the sand: Impossibility of strong watermarking for generative models," in *International Conference on Machine Learning (ICML)*, https://proceedings.mlr.press/v235/zhang24o.html, 2024.

[94] A. Müller, D. Lukovnikov, J. Thietke, A. Fischer, and E. Quiring, "Black-box forgery attacks on semantic watermarks for diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, https://arxiv.org/abs/2412.03283, 2025.

[95] H. Farid, "An overview of perceptual hashing," *Journal of Online Trust and Safety*, 2021, https://doi.org/10.54501/jots.v1i1.24.

[96] Google, *How Content ID works - YouTube Help*, https://support.google.com/youtube/answer/2797370, Accessed April 8, 2025.

[97] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, https://doi.org/10.1109/CVPR42600.2020.00872, 2020.

[98] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, https://doi.org/10.1109/CVPR52729.2023.02345, 2023.

[99] B. Chen, J. Zeng, J. Yang, and R. Yang, "DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images," in *International Conference on Machine Learning (ICML)*, https://openreview.net/forum?id=oRLwyayrh1, 2024.

[100] J. Ricker, D. Lukovnikov, and A. Fischer, "AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, https://doi.org/10.1109/CVPR52733.2024.00872, 2024.

[101] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN generated images easy to detect? A critical analysis of the state-of-the-art," in *IEEE International Conference on Multimedia and Expo (ICME)*, https://doi.org/10.1109/ICME51207.2021.9428429, 2021.

[102] M. Saberi, V. S. Sadasivan, K. Rezaei, *et al.*, "Robustness of AI-image detectors: Fundamental limits and practical attacks," in *International Conference on Learning Representations (ICLR)*, https://openreview.net/forum?id=dLoAdIKENc, 2024.

[103] S. M. Abdullah, A. Cheruvu, S. Kanchi, *et al.*, "An analysis of recent advances in deepfake image detection in an evolving threat landscape," in *IEEE Symposium on Security and*

Privacy (S&P), https://doi.ieeecomputersociety.org/10.1109/SP54263.2024.00194, 2024.

[104] R Core Team, *R: A language and environment for statistical computing*, https://www.R-project.org/, R Foundation for Statistical Computing, Vienna, Austria, 2024.

[105] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, 2015, https://doi.org/10.18637/jss.v067.i01.

[106] H. Singmann, B. Bolker, J. Westfall, F. Aust, and M. S. Ben-Shachar, *Afex: Analysis of factorial experiments*, R package version 1.4-1, https://CRAN.R-project.org/package=afex, 2024.

[107] R. V. Lenth, *Emmeans: Estimated marginal means, aka least-squares means*, R package version 1.11.0, https://CRAN.R-project.org/package=emmeans, 2025.

[108] D. Lüdecke, *Sjplot: Data visualization for statistics in social science*, R package version 2.8.17, https://CRAN.R-project.org/package=sjPlot, 2024.

[109] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. 2016, https://doi.org/10.1007/978-3-319-24277-4.

A Additional Background on Labeling Mechanisms

AI labels displayed on social media platforms can be either added by the users themselves (self-disclosure) or through indirect disclosure methods [83]. In the following, we briefly explain the currently most relevant mechanisms.

Metadata. A straightforward approach is to proactively embed information into a file's metadata. It can be either unsigned, like the IPTC Photo Metadata Standard [31], or signed, like the C2PA standard [30]. The latter uses cryptographic signatures to establish a so-called hard binding between content and its metadata. Notably, these standards are not mainly meant to disclose content as AI-generated, but rather to provide detailed provenance information about the origin, authorship, and editing history of authentic photos. The major disadvantage of metadata is that it can be removed, either intentionally, to conceal AIGC, or automatically because it is stripped when uploaded to non-compliant platforms. Moreover, all parties involved, i.e., providers of GenAI tools, camera manufacturers, developers of editing software, and social media platforms, need to agree on a common standard.

Watermarking. Similar to metadata, watermarking proactively embeds information into the content. However, it is not appended to the file but directly inserted into the content, e.g., the pixels of an image. Visible watermarks are an established means to prevent the unauthorized use of copyrighted material, e.g., stock photos. It has also been shown that deep neural networks can add invisible watermarks to images to embed information [84], [85]. With the emergence of GenAI, novel approaches [86]–[90] perform watermarking during the generation process, such that all produced content can be detected and attributed to the respective model. While these watermarks cannot be simply removed and also have been shown to be robust against minor edits, recent works demonstrate that an attacker can still strip or spoof them [91]–[94].

Fingerprinting. In contrast to metadata and watermarking, fingerprinting does not add information to the content itself. Instead, when an image is created, the model provider computes and stores a hash in a database. This hash can be either cryptographic or perceptual [95], the latter making the fingerprint robust against minor variations and edits. To retrieve the information for a given piece of content, the hash is computed and looked up in the database. A similar system is already used within YouTube's Content ID system to flag copyrighted content [96]. Besides requiring a trusted database, the use of perceptual hashing can cause the hashes of two similar pieces of content to be similar, preventing unique identification.

Detection. The previously mentioned approaches are all proactive, meaning that they rely on the participation of all relevant parties, e.g., camera manufacturers or generative AI providers. Malicious actors attempting to spread AI-generated disinformation will naturally try to circumvent these measures, e.g., by using their own generative model. Passive detection methods do not require proactive measures but exploit artifacts of the generation process to distinguish real from synthetic content [97]–[100]. However, the accuracy of these detectors often deteriorates when content is (adversarially) perturbed [101]–[103].

B Focus Groups

B.1 Study Information and Consent Form

- **This study's purpose** is to produce a scientific publication using anonymized data from the information you provide, with possible anonymous quotes from the focus group.
- **Eligibility** is open to individuals (1.) over the age of 18 (2.) who are active on social media platforms and (3.) are aware of the existence of AI-generated content.
- A subsequent **focus group will be video recorded and transcribed** (converted to text) for analysis purposes by in-house automated transcription software or a GDPR-compliant external service.
- **Personal or project-related information** (e.g. your name) **will be removed** from the transcription (anonymized). We may only publish aggregated data or short quotes in our subsequent publication, without any traceability to you. We will delete the original record of the focus group after its transcription.
- **All study data will be hosted in a secure cloud or on internal servers** accessible only by project members, except in the case of external transcription.
- **Transcribed and anonymized data are kept for up to 10 years** in the spirit of good scientific practice, e.g. if questions about details arise later.
- We expect the focus groups to take up **roughly 90 minutes** of your time. We offer a **compensation of £23.75** for all participants that attended the focus group.
- Your **participation is voluntary**. You may stop participating at any time by closing the browser window or the program to withdraw from the survey. During the focus group, you may decide to drop out of it at any time. If you decide to withdraw your participation, we will delete your contribution to the focus group from the transcript.
- The **risks to your participation** in this study are those associated with basic computer tasks, including boredom, fatigue, mild stress, or breach of confidentiality. The **benefits to you are your** compensation and the learning experience from participating in a research study. The **benefit to society** is the contribution to scientific knowledge.
- For any questions about this research, you may contact: [blinded for review]

By signing this consent form, I am affirming that...

- I have read and understand the above information.
- I am 18+ and eligible to participate in this study.
- I am comfortable using the English language to participate in this study.
- I have chosen to participate in this study. I understand that I may stop participating at any time without penalty.
- I am aware that I may revoke my consent at any time by contacting the research team.
- I am aware that a follow-up focus group will be video recorded.

B.2 Guide

Here, we provide the questions we used to guide our semi-structured focus group discussions. Before the actual questions, the interviewers presented themselves and the purpose of the study, and participants were informed about how focus groups work and asked for consent regarding the use of their data. During the focus group, participants saw a slide deck showing the current topic and exemplary images.

Part A: Generative AI and Risks. Generative AI systems are able to generate new content based on user input. A well-known example is ChatGPT, which can understand and answer questions from users. Another application is the generation of media, e.g., images, using simple descriptions in text form (e.g., the description "A photo of a dog."). The images are generated within a short amount of time and the users do not need to have any prior knowledge about image creation. The images generated in this way can appear very realistic and are increasingly difficult to distinguish from real media.

- **Q1.** I will now ask everybody in turn, have you ever encountered AI-generated images?
  - **Q1a.** What was it?
  - **Q1b.** Where did you encounter it, for example, on which social media platform?
- **Q2.** How did you recognize that the image was AI-generated?
  - **Q2a.** How easy or hard did you find it to recognize that the image was AI-generated?

AI-generated images can be used for many different applications, e.g., as educational content, for better illustration, or even for artistic purposes. However, AI images can also be problematic.

- **Q3.** Can you think of any problems of AI-generated images?

Part B: Opinion and Expectations of AI Labeling. As there is already misinformation that is created with AI today and fears are that this problem will continue to grow, efforts are being made to stop AI-generated disinformation. As one measure, various websites have started to identify and label images created with AI. Politically, this labeling is enforced, for example, in the Digital Service Act of the EU for very large online platforms.

- **Q1.** Have you ever heard about AI labels or even encountered them yourself?
- **Q2.** What is your opinion towards such labels?
  - **Q2a.** Do you find the idea of labeling AI-generated images helpful or not?
  - **Q2b.** If you see such a label on an image, what would be your first thought/impression?
  - **Q2c.** Do you think that such labels could also protect against disinformation?
- **Q3.** Would you say that all AI-generated images should be labeled or only specific ones, like images that are misleading or could falsely appear to be authentic?
- **Q4.** Would you also label images that are edited using AI? One example is that AI filters are used to enhance the image or that the image background is adjusted, like removing a person.
- **Q5.** Who should be responsible for making such AI labeling rules and enforcing them?

Part C: Problems of AI Labeling.

- **Q1.** How do you think that mechanisms to label AI generated images look like?

We will now present three methods of identifying AI images and would like to hear your opinion on them. The simplest way to label AI-generated content on social media is self-disclosure. This means that when uploading something, the user is responsible to mark their content if it was created using AI. Another technique is to automatically detect AI-generated images. These detectors typically also use AI and predict a score denoting how likely an image is AI-generated. The social media platform could apply this detector to all uploaded images and put a label on those that are found to be AI-generated. A third option is to use metadata, which is embedded into an image when it is created. If you use an online service to generate an image, the name of this service and some additional information will be linked to the image file. Once you upload it, the platform can read this data and display the corresponding label.

- **Q2.** What do you think about these approaches?
  - **Q2a.** What do you think are advantages and disadvantages of each approach?
  - **Q2b.** I will now ask everybody in turn, just your gut feeling, which of these approaches do you like the most?
  - **Q2c.** Is there any approach where you would not trust the labels?

We will now talk about the problems each of the three approaches have. With self-disclosure, people could just not indicate that they used AI to create an image, either intentionally or because they took an image from somewhere else and simply don't know. People could also wrongfully say they used AI, reducing trust in the label. Detectors can make wrong predictions (e.g., due to image processing or unseen generative models). This can cause false negatives (AI-generated image is not labeled) and false positives (real image is labeled as AI). The main problem of metadata is that it can be removed intentionally (e.g., by taking a screenshot) or unintentionally (metadata is usually stripped during upload to social media platforms). Moreover, this approach only works if the providers of generative AI tools support the metadata. The approach can also be bypassed by using your own generative model.

- **Q3.** Were you particularly surprised by any of the problems mentioned for the approaches?
- **Q4.** How do you rate the consequences if AI images are wrongfully not labeled?
- **Q5.** How do you rate the consequences if authentic images are mislabeled as AI-generated content?
- **Q6.** What do you consider worse, AI-generated images that are not labeled or authentic images that are mislabeled as AI-generated?
- **Q7.** How does mislabeling affect your opinion and trust in the label?
- **Q8.** Has your opinion towards labeling changed since the start of the focus group after hearing about concrete strategies to mark or detect AI images?
- **Q9.** Is there still anything related to the topic of AI labeling that anyone would like to share with us, maybe something that we forgot to ask?

B.3 Codebook

- A1 Experience of AI
  - A1 Social media
  - A1 Ads
  - A1 Creation of AI
  - A1 News
  - A1 Other websites
  - A1 Seldom/ No experience
- A2 Recognition of AI
  - A2 Context cues
  - A2 Recognition depdends on attention
  - A2 Recognition depends on creator
  - A2 Recognition depends on picture
  - A2 Recognition is easy
  - A2 Recognition is hard
  - A2 Recognition via label
  - A2 Software
  - A2 Visual recognition
- A3 Problems of AI content
  - A3 AI bias
  - A3 Crime (blackmailing, deep porn, scamming etc.)
  - A3 Deception (of skills)
  - A3 Flooding
  - A3 Forged evidence
  - A3 Mis-/Disinformation
  - A3 Unrealistic standarts (beauty, good pics)
  - A3 Availability to Everyone/Traceability
  - A3 Bots
  - A3 Copyright/Privacy issues
- B1 Encountering of AI labels
  - B1 Heard about labels (not encountered)
  - B1 News
  - B1 No encountering of labels
  - B1 Social media
  - B1 Studies
  - B1 Used AI label
- B2 Opinion towards AI labels
  - B2 Helpful but limited
  - B2 Helpful/Positive („they are great“)
  - B2 Labeled content would be perceived negatively
  - B2 Labeled content would be perceived positively
  - B2 Labels could raise acceptance for using AI
  - B2 Should be mandatory
  - B2 Appealing system is important
  - B2 Helpful for specific content
  - B2 Helpful in preventing misinformation
  - B2 Helpful in preventing scams
  - B2 Needed in the long run
  - B2 Spreads awareness about AI content
  - B2 Unsure if perception would change/ Other perception
- B3 AI images that should be labeled
  - B3 Labeling of partly AI-generated images
    - B3 Difficult to decide on labeling rules/ Gray Area
    - B3 Not necessary to label minor AI manipulations (e.g. filters)
    - B3 Risk that labeling gets more complex
    - B3 All partial AI-gen images need label
    - B3 Depends on Content
    - B3 Different label
  - B3 Labeling of completely AI-generated images
    - B3 Difficulty to judge problems of AI images
    - B3 Labels for all AI-gen images
    - B3 Labels for contentious AI images
    - B3 Labels for images containing humans
  - B3 Decision Making/Enforcement
    - B3 Central Organization
    - B3 Community
    - B3 Consistency
    - B3 Creator
    - B3 Government/Law
    - B3 Other
    - B3 Platform
    - B3 Provider of AI
- B4 Problems of AI labels (overarching)
  - B4 Big/complex problem, Standarization is hard
  - B4 Mislabeling could be a problem
  - B4 Overreliance
  - B4 Power of Platform
- C1 Known mechansims
  - C1 AI detection
  - C1 Manual detection
  - C1 Metadata
  - C1 Other
  - C1 Self-disclosure
  - C1 Watermarks
- C5 False positives/false negatives
  - C5 Examples of false negatives
  - C5 Examples of false positives
  - C5 Evaluation
    - C5 Dependent on context/image
    - C5 Equally problematic
    - C5 No lose of trust
    - C5 Source/Credibility of website is important factor
    - C5 More problematic
    - C5 Not problematic
    - C5 Problematic
    - C5 Problematic in the long run
    - C5 Problematic in the short run
  - C5 Consequences
    - C5 Could damage reputation/ trustworthiness
    - C5 Disinformation / leads to questioning of facts
    - C5 Lose of trust if mislabeled images are obvious
    - C5 No consequences (sometimes)
    - C5 Users are getting disturbed/annoyed
    - C5 Users fall for scamming
    - C5 Enables deniability
    - C5 Lose of trust
    - C5 Lose of trust if it happens often
  - C5 Helper Codes: mislabeling

        - C5 Helper Codes: mislabeling: C5 false negative
        - C5 Helper Codes: mislabeling: C5 false positive
        - C5 Helper Codes: mislabeling: C5 general
- M Mechanisms
    - MC HC Mechanism
        - MC HC Mechanism: HC: AI Detection
        - MC HC Mechanism: HC: All/Unspecified
        - MC HC Mechanism: HC: Metadata
        - MC HC Mechanism: HC: Other
        - MC HC Mechanism: HC: Self-Disclosure
        - MC HC Mechanism: HC: Watermarks
    - MC2 Reliability of mechanisms
        - MC2 Changes during time/ AI advances
        - MC2 Mechanism is reliable (for now)
        - C2 Mechanism is not reliable
    - MC3 Preference of mechanisms
        - MC3 Favorite Mechanisms
        - MC3 Combination of mechansims is best
        - MC3 None
    - MC4 Advantages of approach
        - MC4 Independent of user
        - MC4 Scalable
        - MC4 Easy
        - MC4 Independent of AI
    - MC42 Disadvantages of approach
        - MC42 Computing power
        - MC42 Lying/Misunderstanding
        - MC42 Manipulation/Removal
        - MC42 Tool Compliance
        - MC42 Usability
        - MC42 Interpretability
        - MC42 Other
        - MC42 Results dependent on training (model)
- Meta Codes
    - Realization/Surprise
    - Wish
    - Interpretation
    - Interesting quote

C Survey

C.1 Additional Results for Pre-Registered Hypotheses

Confidence. To investigate whether labeling has an influence on the confidence with which people judge claims as true or false, we looked at the confidence ratings, ranging from *very unsure* (1) to *very sure* (4), participants indicated after judging a claim as true or false.

We conducted linear mixed model analyses and fitted models with Group (C, L, M) as fixed effect and the participants' confidence judgment, centered by the grand mean, as the dependent variable. The models had by-subject (i.e., participant) and by-item (i.e., image) random intercepts.

To investigate whether confidence differs between the control group and the two treatment groups (HA4), we compared a model with the first Helmert contrast of Group (including all other main effects) to a simpler model only excluding the first Helmert contrast of Group.

To investigate whether confidence differs between the labeling and the mislabeling group (HB3), we compared a model with the second Helmert contrast of Group (including all other main effects) to a simpler model only excluding the second Helmert contrast of Group.

The model comparisons showed that confidence neither differed between the control group and the two treatment groups nor between the labeling and the mislabeling group (i.e., AIC was lower for both comparison models, $AIC_{Model\_HA4comp}$ = 80703.40 and $AIC_{Model\_HB3comp}$ = 80704.03, $AIC_{Model\_HA4\_HB3}$ = 80711.69).

Response bias. According to the signal detection theory [57], the response bias c indicates participants' tendency to judge claims as true or false regardless of their veracity.

Response bias is calculated as $c = -1 \times \frac{z(H)+z(FA)}{2}$. Positive c scores indicate a conservative tendency, suggesting that people only respond with true when they are very sure about it, negative scores indicate a liberal tendency, suggesting that people respond with true even when they are not entirely sure that the claim is indeed true.

To investigate whether labeling influences the response bias (HA5) and whether response bias differs between the labeling and the mislabeling group (HB4), we calculated a one factorial ANOVA with Group (C, L, M) as the independent and c as the dependent variable and planned Helmert contrasts for the main effect of Group. The overall response bias showed a conservative tendency (c = .47). No pre-registered effect for response bias differences reached statistical significance ($p > .07$).

However, since sensitivity is not controlled for the effect of response bias, we run a post hoc analysis to investigate if the pattern of the Image Type × Veracity × Contrast 1 interaction in accuracy can alternatively be explained by response bias. The Image Type × Contrast 1 interaction contrasts was significant, $t(1351) = 5.43$, $p < .001$, $d = .15$. In both treatment groups the response bias was more conservative for posts containing AIGIs, compared to the control group. For human-made images we find the reversed pattern, that is, a more conservative response behavior in the control group compared to both treatment groups.

C.2 Software

We used the statistical software R Version 4.4.2 for Mac [104] for statistical analyses. For the general linear mixed model and the linear mixed model analyses we used the *lme4* package [105]. For the ANOVA we used the *afex* [106] and *emmeans* package [107], for tables the *sjPlot* package [108] and for the figure the *ggplot2* package [109].

C.3 Study Information and Consent Form

- **This study's purpose** is to produce a scientific publication using anonymized data from the information you provide.
- **Eligibility** is open to individuals (1.) over the age of 18 (2.) who are active on social media platforms and (3.) are aware of the existence of AI-generated content.

- **All study data will be hosted in a secure cloud or on internal servers** accessible only by project members, except in the case of external transcription.
- **Anonymized data are kept for up to 10 years** in the spirit of good scientific practice, e.g., if questions about details arise later.
- We expect the survey to take up **roughly 16 minutes** of your time. **We offer a compensation of £2.86** for all participants completing the survey.
- Y**our participation is voluntary.** You may stop participating at any time by closing the browser window or the program to withdraw from the survey. If you decide to withdraw your participation, we will not utilize your survey answers. You can also opt out of the study after completing the survey by contacting the researchers with your Prolific ID. We will then delete your responses from our dataset.
- The **risks to your participation** in this study involve viewing images or news that are artificial or of a sensitive nature (e.g., content related to politics or violence). The **benefits to you** are your compensation and the learning experience from participating in a research study. The **benefit to society** is the contribution to scientific knowledge.
- For any questions about this research, you may contact: [blinded for review]

By signing this consent form, I am affirming that...

- I have read and understand the above information.
- I am 18+ and eligible to participate in this study.
- I am comfortable using the English language to participate in this study.
- I understand that I may stop participating at any time without penalty.

C.4 Questionnaire

Participants were given the following initial instructions:

> Your task is to identify posts containing false claims that appeared on a social media platform. In the following, you are asked to rate the truthfulness of 26 posts. Each post consists of a short text and an image. The profile image and name of the post's author are anonymized.
> **Control group:** From experience, you know that some posts on the platform contain AI-generated images.
> **Labeling/Mislabeling group:** The platform uses a system to add an "AI-generated" label if an image might be generated using AI.
> Below are two examples of posts: [**control group:** both unlabeled, **labeling/mislabeling group:** one labeled]
> You will only see the post at first, please take a look at it. Shortly after, you will see a question about an associated claim on the right. Please answer the question and indicate how confident you are.
> Clicking on "Next Page" will start the survey.

The following questions were asked for each of the 24 stimuli (see Appendix C.5), plus the two attention checks:

Q1. To the best of your knowledge, <question>? [yes, no]
Q2. How confident are you in your assessment? [very unsure, unsure, sure, very sure]

The following questions were only given to participants in the labeling and mislabeling group.

Instr. You successfully completed the largest part of this survey! We are now interested in your perception of AI labels during the previous task.
Q3. Did you have the impression that the AI labels influenced your decisions in the previous task? [not at all, very little, somewhat, to a great extent]
Instr. The system that the platform uses to add AI labels might not always be 100% correct. It can occur, that images are mislabeled. Mislabeling means that either, an AI-generated image is wrongfully displayed without the "AI-generated" label, or a human-made image is wrongfully displayed with the "AI-generated" label.
Q4. Did you have the impression that, in the previous task, some AI-generated images were not labeled as such? [yes, no, unsure]
Q5. Did you have the impression that, in the previous task, some human-made images were wrongfully labeled as "AI-generated"? [yes, no, unsure]
Instr. We are now interested in your opinion of mislabeling in general.
Q5. Regarding unlabeled AI-generated images, how much do you agree with the following claims? [strongly disagree, disagree, neither agree nor disagree, agree, strongly agree]
  Q5a. Users lose trust in the labeling system if they become aware of such mislabeling.
  Q5b. It is not a problem if such mislabeling only happens once in a while.
Q6. Regarding wrongfully labeled human-made images, how much to you agree with the following claims? [strongly disagree, disagree, neither agree nor disagree, agree, strongly agree]
  Q6a. Users lose trust in the labeling system if they become aware of such mislabeling.
  Q6b. It is not a problem if such mislabeling only happens once in a while.
Q7. Regarding the two types of mislabeling that can occur, which one do you consider worse? [mislabeled (unlabeled) AI-generated images, wrongfully labeled human-made images, they are equally bad, none of them is a problem]
Q8. Would you like to see AI labels (as they were presented in this study) on real-world social media platforms? [definitely no, rather no, neither yes nor no, rather yes, definitely yes]

The last questions were again given to participants in all groups.

Q9. Did you use any tools to rate the truthfulness of the 26 posts, e.g., a search engine or an AI? (Important: There is no right or wrong answer. Your answer does not influence your survey approval.) [yes, no, prefer not to say]
D1. What is your gender? [Woman, Man, Non-binary, Prefer to self-describe, Prefer not to say]
D2. What is your age? [18–24, 25–34, 35–44, 45–54, 55–64, 65+, Prefer not to say]

D3. What is your country of residence? [Dropdown with 249 countries, Other, Prefer not so say]

D4. Which of the following best describes the highest level of formal education that you have completed? [I never completed any formal education, 10th grade or less (e.g. some American high school credit, German Realschule, British GCSE), Secondary school (e.g. American high school, German Realschule or Gymnasium, Spanish or French Baccalaureate, British A-Levels), Trade, technical or vocational training, Some college/university study without earning a degree, Associate degree (A.A., A.S., etc.), Bachelor's degree (B.A., B.S., B.Eng., etc.), Master's degree (M.A., M.S., M.Eng., MBA, etc.), Professional degree (JD, MD, etc.), Other doctoral degrees (Ph.D., Ed.D., etc.), Other, Prefer not to say]

D5. How would you describe your political views? [Very right, Right leaning, Center, Left leaning, Very left, Not interested in politics, Prefer not to say]

C.5 Stimuli

We provide the stimuli used in our survey, separated by subset (human/true, human/false, AI/true, and AI/false) in Figures 4 to 7. Table 4 lists the original and adjusted captions as well as the corresponding questions. We also provide links to fact-checking articles. Figure 8 depicts our attention checks.

ID	Original Caption	Modified Caption	Question	Fact Check URL
Human_True_1	Saving ballots from arson in Vancouver, Washington this morning	A member of the law enforcement saving ballots from arson in Vancouver, Washington	Was there an attempt to save ballots from arson in Vancouver, Washington?	https://www.snopes.com/fact-check/ballots-saved-fire-vancouver-washington/
Human_True_2	Eastern quolls (Dasyurus viverrinus) fluoresce under certain types of UV light	Eastern quolls (a marsupial found in Australia) fluoresce under certain types of UV light	Do eastern quolls glow under certain types of UV light?	https://www.snopes.com/fact-check/fluorescent-marsupial/
Human_True_3	Valencia this morning. A jaw-dropping 343 mm of rain was recorded in just 4 hours yesterday, between 4:30pm and 8:30pm.	Valencia in October 2024. 343 mm of rain was recorded in just 4 hours, causing cars to pile up in the streets.	Did cars pile up in the streets due to heavy rainfall in Valencia?	https://www.snopes.com/fact-check/valencia-spain-flooding-photo/
Human_True_4	Pyongyang's diplomatic community was invited to the opening of the Rungna People's Pleasure Ground. This included the Chargé d'Affaires of the British Embassy, who accepted the invitation to attend. While this was an unusual event, it is vital that we actively engage with the North Korean administration since we work closely with them on a number of humanitarian, cultural and education projects which benefit the people of North Korea.	At the opening of a new theme park, a british diplomat rode in a rollercoaster with Kim Jong Un.	Did Kim Jong Un ride a rollercoaster with a British diplomat?	https://www.motherjones.com/politics/2012/08/kim-jong-un-amusement-park-photo-british-diplomat/
Human_True_5	Former UN Ambassador Nikki Haley signed Israeli artillery shells with the message "Finish Them!" Conflict is no place for stunts. Conflict has rules. Civilians must be protected.	Former UN Ambassador Nikki Haley signed Israeli artillery shells with the message "Finish Them!"	Did former UN Ambassador Nikki Haley sign Israeli artillery shells with the message "Finish Them!"	https://www.snopes.com/fact-check/haley-bombs-finish-them/
Human_True_6	If you're getting annoyed by a Taylor Swift fan just upload this photo and watch them vanish.	Taylor Swift poses with a Swastika.	Did Taylor Swift pose with a man in a swastika-adorned shirt?	https://www.snopes.com/fact-check/taylor-swift-swastika-picture/
Human_False_1	Hyde park this morning, the eco worriers #ExtinctionRebellion have left their plastic rubbish scattered across the park. so much care and concern for the earth is quite touching really!!	Hyde park this morning, members of #ExtinctionRebelion have left their plastic rubbish scattered across the park.	Did members of Extinction Rebellion left Hyde Park full of garbage?	https://www.snopes.com/fact-check/protesters-hyde-park-rubbish/
Human_False_2	There Are a New Animal Species Taking Over at Chernobyl.	There is a new animal species taking over at Chernobyl.	Has a previously unknown species been seen in Chernobyl?	https://www.snopes.com/fact-check/chernobyl-animals/
Human_False_3	MARTIAL LAW DECLARED IN SOUTH KOREA. President Yoon Suk Yeol has announced Emergency Martial Law, with convoys of tanks and helicopters spotted across the capital, Seoul.	President Yoon Suk Yeol has announced Emergency Martial Law, with convoys of tanks and helicopters spotted across the capital, Seoul.	Did military convoys drive through Seoul after martial law was declared?	https://www.snopes.com/fact-check/martial-law-south-korea-photo/
Human_False_4	JUST IN: Italy begins dumping migrants at the door of the Vatican City after Pope Francis said it is a 'sin' to 'reject migrants'.	Italy begins bringing migrants at the door of the Vatican City after Pope Francis said it is a 'sin' to 'reject migrants'.	Did Italy bring migrants to Vatican city?	https://www.reuters.com/fact-check/italy-did-not-transfer-crowd-migrants-vatican-november-2024-12-11/
Human_False_5	Syrian investigative Journalist, Abdul bin Khalid has found the crash site of the plane once carrying former President of Syria, Bashar Al-Assad.	Syrian investigative Journalist, Abdul bin Khalid has found the crash site of the plane once carrying former President of Syria, Bashar Al-Assad.	Did Bashar Al-Assad crash with a plane?	https://www.dw.com/en/fact-check-fakes-surrounding-assads-escape-to-moscow/a-71016174
Human_False_6	P Diddy's mansion in California has been completely consumed by fire.	Sean "Diddy" Combs mansion in California has been completely consumed by fire.	Has Sean "Diddy" Combs mansion in California been consumed by fire?	https://www.reuters.com/fact-check/photo-2014-fire-mislabeled-combs-la-mansion-2025-2025-02-06/
AI_True_1	This is what the French capital city, Paris, looks like. The dream city... now turned into this in reality	The streets of the French capital city, Paris, are filled with garbage after a three-week strike of garbage collectors.	Did Parisian garbage collectors went on strike, causing uncollected garbage littering the streets?	https://factcheck.afp.com/doc.afp.com.33QV2QL
AI_True_2	Rare pink Dolphin spotted in Bohol	Since 1962, only 14 pink bottlenose dolphins have been spotted.	Is there a species of dolphins that is pink?	https://factcheck.afp.com/doc.afp.com.34ZE9BT
AI_True_3	This is the home of a Christian in Los Angeles, California. While the houses around him were destroyed by fire, his house remained untouched. God's promise in Psalm 91:1-6 was fulfilled. You can't imagine how much he cried for joy, knowing he was protected by God. Truly, God is his refuge.	After the wildfires in Los Angeles, California. While the houses around were destroyed by fire, a single house remained untouched.	Did a single house remain untouched while the houses around it were destroyed during the LA wildfires?	https://www.snopes.com/news/2025/01/14/la-fires-home-god-saved/
AI_True_4	LOOK: Picture of about two million young people that attended Mass with Pope Francis in Lisbon! I'm Catholic For Life! #WorldYouthDay2023	About 1.5 million young people attended Mass with Pope Francis in Portugal celebrating World Youth Day!	Did 1.5 million people attend mass with Pope Francis in Portugal celebrating World Youth Day?	https://factcheck.afp.com/doc.afp.com.33R24HY
AI_True_5	This is Beirut tonight this is not self-defense.	Commercial flights landing at Beirut International Airport despite Israeli airstrikes.	Did the airport in Beirut still operate despite airstrikes?	https://www.reuters.com/fact-check/images-aircraft-landings-into-flaming-beirut-airport-are-ai-generated-2024-10-29/
AI_True_6	First Look at Lady Gaga in WEDNESDAY Season 2!	Lady Gaga to appear in 'Wednesday' Season 2.	Is Lady Gaga going to appear in 'Wednesday' Season 2?	https://www.comingsoon.net/guides/news/1894416-wednesday-season-2-lady-gaga-first-look-image-real-fake-ai
AI_False_1	A 57,000 square foot Temu warehouse in China went up in flames today. The total loss of inventory has been estimated to be as high as $56.19 USD.	A 57,000 square foot Temu warehouse in China went up in flames.	Did a Temu warehouse in China go up in flames?	https://www.snopes.com/fact-check/temu-warehouse-fire-china/
AI_False_2	A giant octopus was discovered off the coast of Bali, Indonesia.	A giant octopus was discovered off the coast of Bali, Indonesia.	Was a giant octopus discovered off the coast of Bali, Indonesia?	https://www.snopes.com/fact-check/giant-octopus-indonesian-coast/
AI_False_3	It is reported that Disneyland has been flooded due to Hurricane Milton.	Disneyland has been flooded due to Hurricane Milton.	Has Disneyland been flooded due to Hurricane Milton?	https://factcheck.afp.com/doc.afp.com.36JU2AM
AI_False_4	This is hysterical. The President of Mexico was spotted wearing a 'Make America Mexicana Again.'	The President of Mexico was spotted wearing a 'Make America Mexicana Again' hat.	Was the President of Mexico spotted wearing a 'Make America Mexicana Again' hat?	https://leadstories.com/hoax-alert/2025/02/fact-check-mexican-president-claudia-sheinbaum-did-not-wear-make-america-mexicana-again-hat-it-was-made-using-xs-ai-tool.html
AI_False_5	Image released of the migrant mob that stormed a Birmingham hospital armed with axes and blades.	A migrant mob stormed the Birmingham hospital in England armed with axes and blades.	Did a migrant mob storm a hospital in Birmingham, England?	https://www.reuters.com/fact-check/image-armed-hospital-ambush-is-ai-not-evidence-disorder-uk-2025-02-21/
AI_False_6	Keanu Reeves is playing Bob Marley in a new movie he's shooting in Jamaica.	Keanu Reeves is playing Bob Marley in a new movie he's shooting in Jamaica.	Is Keanu Reeves playing Bob Marley in a new movie?	https://www.snopes.com/fact-check/keanu-reeves-dreads/

Table 4: Overview of our stimuli's original caption, modified caption (used in our survey), corresponding question, and the link to a fact check. Note that in our survey, each question was introduced by "To the best of your knowledge, ...?"

(a) Human_True_1

(b) Human_True_2

(c) Human_True_3

(d) Human_True_4

(e) Human_True_5

(f) Human_True_6

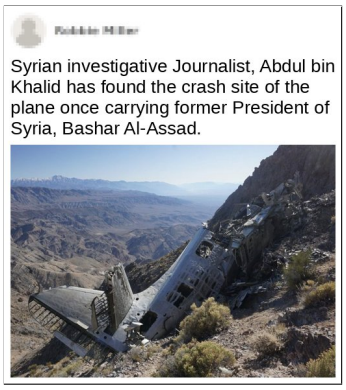
Figure 4: Stimuli in the *human/true* subset

(a) Human_False_1

(b) Human_False_2

(c) Human_False_3

(d) Human_False_4



(e) Human_False_5

(f) Human_False_6

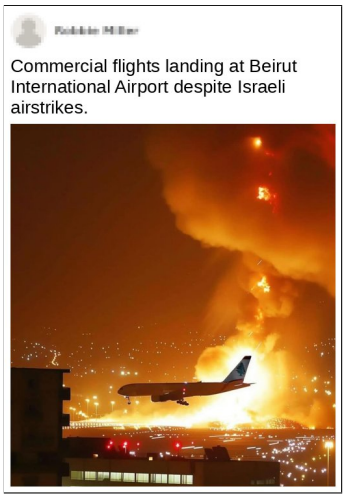
Figure 5: Stimuli in the *human/false* subset.

(a) AI_True_1

(b) AI_True_2

(c) AI_True_3

(d) AI_True_4



(e) AI_True_5

(f) AI_True_6

Figure 6: Stimuli in the *AI/true* subset.

(a) AI_False_1

(b) AI_False_2

(c) AI_False_3

(d) AI_False_4

(e) AI_False_5

(f) AI_False_6

Figure 7: Stimuli in the *AI/false* subset.

(a) First attention check, shown after the tenth post.

(b) Second attention check, shown after the 20th post.

Figure 8: Attention checks.