"It says AI-generated content, but we cannot see it.": Practices & Challenges of Navigating Indicators for AI-Generated Content for Sighted and Blind Individuals

Ayae Ide

AI-Generated content (AIGC) has become increasingly widespread by recent advances in generative models and the easy-to-use tools that have significantly lowered the technical barriers for producing highly realistic audio, images, and videos through simple natural language prompts. In response, platforms are adopting provable provenance by recommending AIGC to be disclosed and signaled to users. However, these indicators may often be missed, especially when they rely solely on visual cues, making them ineffective for users with different sensory abilities. To address the gap, we conducted semi-structured interviews (N=28) with 15 sighted and 13 blind and low-vision (BLV) participants to examine their interactions with AIGC and AI disclosure indicators. Our findings reveal diverse mental models and practices, highlighting different strengths and weaknesses of content-based (e.g., title, description) and AIGC-specific (e.g., AI labels) indicators. While sighted participants leveraged visual and audio cues, BLV participants primarily relied on audio and existing assistive tools, limiting their ability to identify AIGC. Across both groups, they frequently overlooked AIGC-specific indicators and more often interacted with content-based indicators such as title, description, and comments. We uncovered usability challenges stemming from inconsistent indicator placement, unclear (unstructured) metadata, and cognitive overload. These challenges are further amplified for BLV individuals due to the insufficient accessibility of interface elements. We provide practical recommendations and design implications for future AIGC indicators across several dimensions.

CCS Concepts: • **Human-centered computing** → **User studies**.

Additional Key Words and Phrases: AI-generated Content, Deepfakes, Social Media, Mental Model

ACM Reference Format:

Ayae Ide. 2026. "It says AI-generated content, but we cannot see it.": Practices & Challenges of Navigating Indicators for AI-Generated Content for Sighted and Blind Individuals . In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (CSCW '26)*. ACM, New York, NY, USA, 37 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

AI-Generated content (AIGC)[1] has become increasingly widespread by recent advances in generative models and the easy-to-use tools such as ChatGPT, Adobe Firefly, Midjourney, Synthesia, and ElevenLabs. These increasingly public-facing tools have significantly lowered the technical barrier to producing highly realistic text, images, and videos through simple natural language prompts. Historically, prior forms of media alteration, such as photo retouching, airbrushing, or Photoshop-based editing, were often normalized and culturally accepted as part of established media production practices [54, 74]; however, as AIGC has significantly decreased the technical skills and effort needed to

[1]AI-generated content refers to content created or manipulated by algorithms, for example, videos that swap one person's face onto another's face, photorealistic synthetic images, cloning of a person's voice [27, 102, 120, 157]

Author's Contact Information: Anonymous.

Manuscript submitted to ACM

generate synthetic content, a significant amount of AIGC now exists on social media platforms [103, 136]. Futhermore, while AIGC expands expressive and collaborative possibilities [69, 77, 139], it also exhausts the existing frameworks for assessing authorship, credibility, intent, accountability, and trust in digital environments [49, 87, 95, 117, 150] due to its unprecedented scale, speed, and automation. As a result, AIGC is becoming increasingly indistinguishable from human-generated [106, 107], with evident harms enabled by fabricated political messaging [12], AI-driven sexual harassment [59], and targeted financial scams [123], among others.

To combat these risks, there is a large body of literature focused on reliably distinguishing AIGC content from non-AI-originating content. *AIGC detection methods* automatically determine whether content, such as audio, video, and images, has been manipulated either by humans or generative AI [24, 125, 155]; unfortunately, this largely has resulted in a cat-and-mouse game with no end in sight [88]. Consequently, defenses are shifting toward *media provenance*, in which standards, such as C2PA [3, 53, 119], embed cryptographically verifiable metadata alongside media to track the content's origin and modifications.

Once information about AI generation or manipulation is available, it is typically communicated to users via AIGC indicators. These indicators fall into two broad categories: (1) *AIGC-specific Indicators:* Platforms may explicitly label content as AI-generated via visual overlays (e.g., icons), warnings, or standardized tags embedded in the media description [10, 15] uniquely made to convey this information. While standardized within a platform, these indicators often vary between platforms. (2) *Content-based Indicators:* Content creators may voluntarily disclose AI generation through informal signals within their posted media. For instance, creators may use a title, a hashtag, or embedded text in their media to indicate that their content is AIGC; however, because these are informal and non-standardized, they may vary across creators.

However, even with regulatory and technical progress, significant usability challenges persist. Users still encounter AIGC, as many instances go undetected and several are not intentionally deceptive [6]. Prior work suggests that users often conflate provenance metadata with the truthfulness of content and confuse the credibility of the provenance with that of the content itself [51]. Furthermore, issues in indicator usability are often exacerbated for underserved populations within unique accessibility requirements, for instance, Blind and Low-Vision (BLV) users [144]. Historically, BLV users and their interface requirements (e.g., screen reader accessibility, non-visual sensory modalities) are often not considered when designing computer interfaces [134, pg. 24-39], leading to events like 70% of BLV users being unable to access misinformation warnings due to poor accessibility and interface design [131]. This is particularly concerning in sensitive contexts, where deceptive AIGC of public figures can spread misinformation and shape public perception [13, 14, 30]. Thus, while AIGC indicators are becoming common [8, 9, 16], we know little about how well sighted users or BLV users navigate environments saturated with potentially deceptive media, or whether current indicator-based mitigations meaningfully support their needs.

To fill this gap, we follow best practices of inclusive security [144] and universal usability [134, pg. 24-39] and ask for both sighted and BLV stakeholders:

RQ1: How do sighted and BLV individuals conceptualize AI-generated content?
RQ2: What do sighted and BLV individuals anticipate as the benefits and harms of AI-generated content?
RQ3: How do sighted and BLV individuals interpret and engage with AI-generated content indicators?
RQ4: What improvements do sighted and BLV individuals recommend for AI-generated content indicators?

To address the research questions, we conducted semi-structured interviews with 15 sighted and 13 BLV participants to investigate their conceptual understandings of AIGC. To ground participants' reflections in concrete encounters, we

incorporated an interactive screen-sharing session using 12 pre-selected AI-generated content samples that vary by platforms (e.g., YouTube, TikTok, Instagram), formats (e.g., video, audio, image), categories (e.g., Politics, Entertainment, Commercial, Music, Religion) and indicator types (AIGC-specific indicators, content-based indicators). We leveraged these samples, not to be exhaustive, but to serve as probes that allow participants to articulate their understanding and key pain points in informing design recommendations among a set of exemplars AIGC indicators.

Main findings. Our study highlighted mental models among participants constructed based on their current understanding, practices, and risk-benefit trade-offs. These models include conceptions of AIGC as either user-led creation or algorithmic production, approaches to AIGC identification which juxtaposes perceived content reliability with the need for external verification, sensory modality in varying level of AIGC assessment and a risk–benefit framing shaped by differential dependence on AI-mediated infrastructures, in which BLV participants emphasized functional reliance on AI-powered accessibility technologies, while sighted participants more frequently expressed concern about long-term creative degradation and authenticity erosion. We find participants leveraging their own tactics and various cues to discern AIGC, including voice (e.g., tone, pitch, monotone, breathing), content quality (e.g., too polished, too unrealistic, too perfect to be human-made), face and scene movement, distortion, etc, in addition to common AI indicators in the title, description, etc.

Our interactive session with participants highlighted the strengths and weaknesses of both content-based indicators (e.g., title, description, hashtag, comments, and creator watermark) and AIGC-specific indicators (e.g., AI labels). Participants tend to interact more with content-based indicators than the AIGC-specific indicators due to visibility and prior familiarity with that content. Our findings highlight participants' preferences for the future design and development of indicators across several dimensions, including content, time, placement, modality, and creation-platform responsibility.

2 RELATED WORK

2.1 Evolution of AI-generated Content & Deepfake

The rapid progress of generative AI has enabled the development of text, images, audio, and video [73, 122]. While these tools offer creative and automation benefits, they also may lead to the creation and dissemination of deceptive content, such as deepfakes, which can raise political and social concerns [50, 116]. For instance, political generative content may increase distrust in media [39, 140], while can spread disinformation, cause psychological harm, and erode public trust [22, 29]. Non-consensual AI-generated intimate imagery harms reputations [101], while cloned voices are increasingly exploited in scams to impersonate family members [11]. In entertainment and advertising, AI is used to recreate deceased actors and deploy virtual influencers, raising concerns about consent and deception [28, 108]. In music, generative tools intensify copyright disputes over unauthorized voice cloning [127]. People misclassified deepfake videos and overestimated their abilities [84, 138], listeners identified synthetic speech slightly better than chance [99], and neither humans nor AI systems achieved high accuracy in detecting fake images [64]. These findings highlight alternatives in which provenance indicators are used [149]. While prior work has examined detection and labeling [23, 149], less is known about how users interpret and act on provenance indicators in everyday media consumption.

2.2 Emerging Media Provenance and Indicators for AI-Generated Content

With the proliferation of AI-generated content, much prior work has developed techniques to assess content authenticity. The core technology of media provenance comprises three categories: metadata, fingerprinting, and watermarking [41]. In the following, we outline recent advancements and representative techniques in each category.

AIGC Detection. Fingerprinting techniques extract distinctive features from media content itself, such as patterns or noise artifacts. Prior work shows that GANs and diffusion models leave model-specific fingerprints that enable attribution [55, 94, 152, 154], GAN artifacts [154], 3D head-pose inconsistencies [152], and hybrid camera–synthetic fingerprints [91] that can be used to identify and distinguish AI-generated or manipulated images. For audio, recent studies have proposed detecting fake audio by identifying unique fingerprints left by the tools that create it [121, 151].

AIGC Provenance. Broadly, provenance systems for AI-generated content differ based on where provenance information is stored and how it is verified. Existing approaches fall into three categories, including **embedded provenance** which binds origin and editing histories directly to media files. One prominent industry standard is Coalition for Content Provenance and Authenticity (C2PA) [4], which uses cryptographic hashes and signatures to ensure that any alteration to asset provenance can be detected. Its "content credentials" are being integrated into cameras, creative tools, and generative models such as Adobe Firefly and OpenAI DALL·E, with adoption supported by the Content Authenticity Initiative (CAI) [17]. Another category is **ledger-based provenance** where signed provenance records are maintained in external infrastructures. For example, AMP [48] proposed using distributed ledgers to track the origin and authenticity of AI-generated media, while VerITAS [45] extends C2PA with zero-knowledge proofs to verify that only permissible edits were applied. A complementary class of provenance techniques relies on **signal-based approaches**, such as watermarking that embeds invisible signals into media to verify origin or detect tampering. Traditional watermarking methods, such as spread-spectrum watermarking [43] and quantization index modulation [38], laid the foundation for robust techniques resilient to compression and minor edits. Recent work embeds watermarks directly into generative models [66, 75, 76, 96, 146], including diffusion-based approaches that preserve quality and resist advanced edits.

Human Perception of AIGC Indicators. Recent work has increasingly discussed how to label AIGC and how these indicators influence users' judgments and interpretation of content [37, 57, 143, 148, 149]. Prior work [148] categorized the labeling of AIGC into two strategies: *"process-based"* and *"impact-based"*, each focusing on technical processes of content generation and potential harm. More recently, they found that different labeling strategies for AIGC may affect viewers' likelihood of engagement [149]. In a study examining viewers and creators' perceptions of content authenticity and AI labeling solutions, the authors emphasized that labels must align with viewers' mental models and information needs [37]. Research shows that labels shape users' beliefs about content, with label design affecting users' trust in the label [57]. However, there remains a lack of evidence on how BLV users experience the process of establishing provenance. Recent work on AI virtual assistance technologies underscored the importance of valuing blind epistemologies, recognizing that blind users develop unique verification strategies to AI outputs [23].

2.3 BLV Users Interaction with AI- Generated Content

Previous research indicates that BLV individuals interact with visual and audio media aimed at enhancing self-expression and social involvement [19, 33, 62, 71, 78, 142]. Nevertheless, there is little focus on how they interact with the content's authenticity.

Challenges of BLVs in Navigating AIGC. Recent studies have shown that most users struggle to detect AI-generated voices [32, 72, 110], and often users rely on linguistic cues and intuition to classify but are often misled by preconceptions about generated audio, such as assuming accents or background sounds indicate human speech [145]. Blind and low-vision (BLV) individuals, in particular, face challenges in detecting audio deepfakes, including manipulated conversational speech and impersonation attacks, where studies show the audio deepfake discernment accuracy for BLV users is at approximately 50%, and may decrease further as generative AI continues to rapidly improve [130]. Han et al. [68] observed differences between sighted and BLV users in detection strategies and accuracy

when identifying real and spoofed audio. While sighted users can rely on visual cues, such as forensic labels [133], BLV users must rely on alternative cues, such as alt text, which are often inaccessible or incomplete [18, 44]. Although assistive technologies and automated image descriptions offer basic content information [98, 128], they rarely include provenance details.

Challenges of BLV with AIGC Indicators. Inaccessibility for BLV users is not a new phenomenon introduced by AIGC indicators, but rather traces historical challenges in security warning design, including phishing warnings, misinformation labels, HTTPS encryption information, and SSL certificates [82, 112, 131, 156]. Empirical evidence demonstrates that the effectiveness of warning design often depends on its implementation, given existing accessibility guidelines [111]. Thus, a well-intended warning can still be ineffective if it is not delivered through accessible channels and is incompatible with assistive technologies. Early work on automatic alt-text for social networks is directly situated within media usage and interaction patterns [61, 97]. Particularly, BLV users experience an underserved interpretation of visual descriptions on mainstream platforms [70]. Current AI-generated content indicators share similar design challenges [5, 6, 16, 34], relying largely on visual watermarks or poorly surfaced textual labels with new standards increasingly embedding provenance metadata with visually oriented cues such as the C2PA "CR" logo [7]. Drawing on literature [82, 131], design choices can render AI indicators effectively non-actionable, where information exists technically but not functionally in practice.

2.4 Platform Policy on AI-Generated Content

To understand the current norms and expectations surrounding AI-generated content disclosures, we examined platform policies across major media platforms where AI-generated content is prevalent and widely consumed by everyday users. We outline the current governance of labeling practices and the enforcement of AI content disclosures (Table 4) across platforms, which inform our study design (section 3).

YouTube. YouTube requires creators to disclose with a label when content is generated or significantly altered by AI, particularly if it involves *"the likeness of a realistic person, alters footage of real events or places, or generates realistic scenes"* [15]. YouTube's Creator Studio facilitates this disclosure if the content is AI-generated. With their current mandates, "creators who consistently choose not to disclose this information may be subject to content removal, suspension from the YouTube Partner Program, or other penalties" [10]. Furthermore, YouTube requires content to comply with broader community guidelines that prohibit impersonation, harmful content, scams, and misinformation. Additionally, viewers can submit privacy requests to remove synthetic content that simulates identifiable individuals, including face or voice likenesses [10].

TikTok. TikTok's Edited Media and AI-Generated Content' policy similarly encourages creators to [1] to disclose if their content is AI-generated, with synthetic media replicating realistic-looking images, audio, and videos falling under misinformation, subject to removal under its community guidelines. TikTok supplements creator labeling with Content Credentials from the Coalition for Content Provenance and Authenticity (C2PA) [1]. In this case, once labeled either manually or automatically, the AI tag is permanent.

Meta. Instagram also shares the goal to establish "transparency and trust" [6]. To support this, Meta identifies AI-generated content if a creator uses a Meta AI tool or if the content includes detectable AI signals. This means that after using an AI generator, the image may add a signal showing that *"content was modified using AI"* [6] Unlike YouTube and TikTok, Instagram doesn't require the creator to label images with the AI info label because the AI info label is automatically applied if Meta detects those signals. However, this approach is technically limited, as Meta acknowledges

its inability to detect AI-generated audio or video content that lacks such signals, especially from third-party tools [8]. While Instagram notes that *"there may be penalties if you do not label content as required"*[6], there is a lack of clarity on the policy's effectiveness. Moreover, Instagram states that AI-generated content will only be removed in high-risk scenarios, given that generative AI is becoming a mainstream tool for creative expression [34] t

3 METHOD

We conducted semi-structured interviews to understand the practices and challenges of sighted and BLV users with AI-generated content across different platforms. Our main goal is to ① construct the mental models people hold of AI-generated content, and ② understand how people interact with existing indicators in navigating AI-generated content. The study was approved by the Institutional Review Board (IRB). The interviews took place remotely via Zoom. Each interview session took approximately one hour.

Why Include Both Sighted and BLV Perspectives? Our goal is to improve the usability of the AIGC indicator for all users. To do so, we follow best recommended practices in inclusive security [144] and universal usability [134, pg. 24-39] and, like other works [85], not only include recruits from the general population, but also intentionally include BLV persons, a marginalized, minority group that is often not included in general research. We do this for two reasons: *First*, BLV persons are not only active users of social media, but before our study we suspected that current indicators may differ in usability for this population; indeed, we find and highlight such results. *Second*, designs that are universally usable, and thus informed by the inclusion of minority populations, often benefit all users [134, pg. 24-39] (e.g., curb cuts in sidewalks not only benefit). For these reasons, we include both sighted and BLV users in our study.

3.1 Study Protocol

The session is divided into three sections (in Appendix A). In the **first section**, we sought participants' understandings of AI-generated content (AIGC). For instance, we asked - *"How would you define AI-generated content?"* We then asked participants about their experiences encountering AIGC and how they identify it. While describing how they identified AIGC, we avoided any technical jargon or terminology around "indicators"; however, once described by the participant, we then referred to those signals as "indicators" for the rest of the interview. To further investigate how their perception and lived experience with AIGC shape their trust, benefits, and concerns towards AIGC across diverse media platforms, we asked *"Have you noticed any positive aspects of AI-generated content? Have you observed any harmful uses of AI-generated content?"* For both groups, we asked open-ended questions to elicit participants' own framings of potential harms, positive aspects, and risks related to AIGC. When BLV participants described initial perceived risks or benefits, we followed up by asking whether, based on their experience, AI-generated content might present different or additional risks for people with visual impairments.

In the **second section**, we investigated participants' navigation practices through an interactive screen-sharing think-aloud session on their desktop device. [2] We asked participants to engage with pre-selected AIGC (e.g., images, videos, and audio) across multiple platforms such as YouTube, TikTok, and Instagram (section 3.2 includes content selection criteria). We informed participants that the interview focused on AI-generated content; however, we did not inform if particular items were AI-generated. We designed this session to observe not only how participants identified and interpreted AIGC but also how their behaviors and judgments unfolded during active engagement. We designed this session to examine how participants identified and interpreted AIGC during active engagement. We instruct them

[2]While one participant (Q9) unexpectedly joined the interview via a mobile device (due to lack of access to a computer), this response was otherwise successfully complete and thus kept in this study. Future work should more fully investigate mobile-device interfaces.

to visit the AIGC URLs on various platforms (Appendix Table 6 and Figure 7) To ensure diversity, we curated 12 AIGC examples spanning platforms (YouTube, TikTok, Instagram), formats (video, audio, image), categories (e.g., politics, entertainment, commercial, music, religion), and indicator types. The examples were organized into three groups of four, each covering all platforms and varied categories (G1: 1,4,5,9; G2: 2,6,8,11; G3: 3,7,10,12). Each participant was assigned one group, viewed all four stimuli in randomized order. We used the samples as probes rather than an exhaustive set. During the session, we asked questions to observe their processes and challenges navigating AIGC and its indicators.

(1) **How Users Detect AI-generated Content.** Example questions: *"What are your initial impressions of the video?", "How did you determine this video/image/audio might be AI-generated or authentic?"*
(2) **What Users Think of the Existing AI-generated Content Indicators.** Example questions: *"When you see [this indicator (the cue participants mentioned)], what's the first thing that comes to mind?", "Do you feel [these indicators (the cues participants mentioned)] help you make sense of whether a video/image/audio is AI-generated or real?"*
(3) **How AI-generated Content Indicators Can Be Improved.** Example questions: *"Was it clear where [the indicators (the cues participants mentioned)] were located? Where would you place [the indicators (the cues participants mentioned)] to make them more noticeable?", "When do you think is the best time to show [the indicators (the cues participants mentioned)]?"*

When participants didn't explicitly interact with the indicators or explain their thoughts on the indicators, or shared their judgments solely on content (e.g., speakers' tone and pitch, content types, visual cues), we nudged them to check the AI indicators (e.g., descriptions, comments, AI labels) and encouraged them to review those indicators. For instance, we prompted them to check on-screen information (e.g., title, description) and asked them to describe what they found (e.g., "Could you go to the description? What information is provided there?"). (interview protocol in Appendix A.)

In **the third section,** we explored how participants would like AIGC indicators to be improved based on their needs and experiences. We began with open-ended questions to capture their initial thoughts, then guided the discussion using themes identified during pilot sessions. During the pilot sessions, participants reported challenges across several design dimensions, including placement, timing, and modality, which motivated us to include a set of questions focused on design improvements. For example, when a participant identified challenges related to indicator placement, we followed up with questions to elicit design-improvement suggestions within that dimension. For screen reader users, we avoided the phrase *"where on the screen"* and asked how they encountered the information and how noticeable or discoverable it was with their current accessibility channel. From this, we collect participants' (1) interview responses, (2) observational notes taken during the interview, and (3) a screen recording of participants' interaction with AIGC.

3.2 AI-generated Content Selection

Our goal is to capture the diversity of the current AI-generated content to investigate user perceptions across a range of contexts, modalities, and disclosure practices, building on prevalent use cases from prior work [36, 40, 83, 86, 100, 105, 115, 118, 153]. To ensure the diversity and relevance of stimuli, we selected 12 AI-generated content samples that vary by platforms (e.g., YouTube, TikTok, Instagram), formats (e.g., video, audio, image), categories (e.g., Politics, Entertainment, Commercial, Music, Religion) and AI indicators types (content-based indicators, AIGC-specific AI labels), as illustrated in Table 6[3] This selection was intended to elicit difficulties that would likely be encountered by all users

[3]We excluded X due to differences in mechanism design, as it focuses mainly on account-level AI disclosure [2], with content-level fact-checking handled via Community Notes [135]. Our pilot also showed that a few participants (no BLV participants) used X

Platform	Content-based Indicators					AIGC-specific AI Indicators		
	Title	Description	Hashtag	Comments	Watermarks	Single-line	Description Info	Rotating Label
YouTube	◗	◗	◗	◗	◗	●	●	❍
TikTok	❍	◗	◗	◗	◗	●	❍	❍
Instagram	❍	◗	◗	◗	◗	❍	●	●

Table 1. AI-Generated Content Indicators. We conducted the experiment across platforms where ●= platform-required, ◗= available, but not required, ❍=not supported

(e.g., content knowledge) as well as users with varying vision-ability (e.g., reliance on visual inconsistencies vs. audio cues). Our team carefully reviewed multiple media samples and discussed the relevance of each item for sighted and BLV participants, as well as the anticipated difficulties for each group. This process was conducted through a group discussion among researchers with expertise in HCI, security/privacy, and deepfake/ML. We reached our conclusions through consensus and selected a final set with the highest relevance while ensuring diversity. The outcome of this discussion regarding the final set of samples is summarized in Table 7. Each media item contains several AIGC indicators. We classify them into two categories (Table 1)

(a) Content-based Indicators refer to AI-related information that appears within existing post fields, such as titles, descriptions, hashtags, or comments. These fields are not explicitly designed for AI disclosure, but creators and users can voluntarily include such information there.

(b) AIGC-specific Indicators are explicitly designed for AI disclosure. Platforms may explicitly label content as AI-generated via visual overlays (e.g., icons), warnings, or standardized tags embedded in the media description [10, 15] uniquely made to convey this information.

This classification reflects both the diversity of disclosure practices across platforms and the varying levels of visibility and user effort required to access provenance information. By distinguishing between organic indicators embedded in general post fields and explicit labels designed for AI disclosure, we aim to capture the range of user experiences with AIGC interventions.

3.3 Recruitment and Participants

We recruited participants through (1) mailing list of National Federation of the Blind[4] to recruit BLV participants; (2) online platforms (Prolific) to recruit sighted participants. We invited respondents to our study if they met the selection criteria: a) 18 years or older and b) living in the US. Prior to each interview, participants were informed that the session would involve screen and audio recording. Only participants who agreed to the consent form were invited to the interview. Participation was voluntary, and participants could withdraw at any time. To reduce the risk of potential harm, researchers examined all video materials included in the study to ensure that they did not contain any harmful content. Each participant received a $20 Amazon e-gift card upon completing an hour-long interview. We continued recruitment until we observed conceptual saturation [56, 113] in our interviews (e.g., when additional interviews yielded no new insights). All participants have previously encountered or interacted with AIGC. Among 28 participants, 15 are sighted, seven have some light perception, two are legally blind, and four are completely blind. Table 5 presents the demographics of our participants. We refer to BLV participants as P1,. . . , P13 and sighted participants as Q1,. . . , Q15. Our participant pool reflected a wide spectrum of demographics as follows. The majority of participants (39%) were in the 25–34 age range, followed by 25% in the 35–44 range, and 4% in the 45–54 range. A few participants were aged 18–24 (11%), and only two were over 65. Participants had diverse educational backgrounds: 43% held a postgraduate degree, 32% held a bachelor's degree, and the remainder had associate degrees, vocational training, or high school

[4]NFB: https://nfb.org/

education. About 29% of participants reported having an AI/ML background. All participants had prior exposure to AIGC, with the majority (56%) having encountered all four types: text, image, audio, and video.

3.4 Data Analysis

We analyze participant responses, screen recordings, and our notes from the interview using thematic analysis [35]. Specifically, we first applied a combination of inductive and deductive coding to our recorded data [52]. For instance, *deductive* coding focused on expected a priori, such as audio cues (e.g., voice tone) and visual cues (e.g., facial or body movements) for what content indicator was mentioned for detecting AIGC, whereas *inductive* coding captured unanticipated challenges, how experiences differed or conflicted between visual and nonvisual interaction practices, including incompatibilities between AIGC indicators and screen reader technologies. Using this hybrid strategy, we began with an initial codebook derived from our deductive codes, which described 6 interviews. Using this codebook, we randomly selected 3 interviews and two coders independently coded each, resolved disagreements, and introduced new codes to each other's codebooks. Furthermore, the entire research team held regular meetings to discuss the codes under deliberation and reach a consensus. This was repeated for 3 rounds, with 3 interviews per round, until the codebook stabilized (see Table 8). Using this codebook, a single-coder then independently coded the rest of dataset (10 interviews). Upon coding, we then perform a reflexive thematic analysis, constructing themes and insights from the experiences participants had with AIGC indicators; the resulting themes are reported in our findings.

3.5 Positionality

Our team strongly agrees with community-originating calls [144] for the promotion of universal usability and participatory design principles [134, pg. 24-39]; this heavily informs our study design. Given decades of long-standing historic inequities around multi-media accessibility for BLV users in software and security-specific design [144], and calls for explicit inclusion among the community [144], we explicitly include them as stakeholders in our assessment to promote inclusive access. Furthermore, given the inconsistent design AIGC indicators, we suspected that screenreader accessibility may prove difficult; however, to the best of our ability, we did not let this influence our interpretation or analysis; indeed, when reported in our work, BLV users who were not made aware of the suspicions, note the difficulty of existing indicators which drives our analysis, and to the best of our ability, we transparently present in our results. Our team also has over half a decade of experience conducting research on inclusive security and privacy with the BLV community, as well as on user perceptions of AIG media. Notably, however, our team does not contain any individuals who identify as part of the BLV community, and as such, we acknowledge the limitations in our perspective.

3.6 Ethics and Participant Protections

To mitigate potential harm, researchers reviewed all video materials included in the study to ensure they did not contain harmful content. Our IRB submission included the final study protocol, all URLs used in this study, and the study scripts to assess risks. Prior to the interview, we provided a brief of the study procedures, including how data would be collected, stored, and retained in compliance with Institutional Review Board (IRB) guidelines. We also explained the potential risks of participation, which are minimal, as they are typically encountered during everyday internet use. We believe the potential benefits of informing safer, more accessible AIGC disclosure design outweigh the minimal risks, consistent with prior literature on user exposure to targeted ads for informing privacy-setting design [31, 90, 132]. Prior to each interview, participants were informed that the session would involve a screen recording and would explore the pre-selected content. We instructed participants that participation was voluntary and that they could withdraw at any

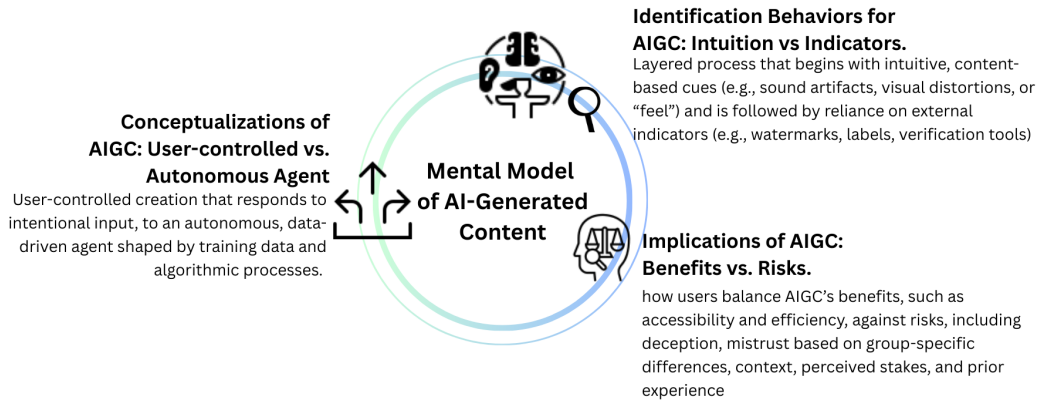

Fig. 1. Participants' Mental Model for AI-generated Content - three dimensions derived from our qualitative results.

time. Finally, after the study, we debriefed on the selected AI-generated content used. We also acknowledge that IRB processes may not keep pace with evolving media. Within the scope of existing standards, we took feasible steps to conduct this research responsibly. We view respect for persons and a robust informed consent process as foundational principles guiding this type of study.

3.7 Limitations

While we find clear instances of usability issues in current AIGC indicators, there are several limitations to our findings. *First*, our stimuli were not constructed to span all combinations of platform labeling and comment signals; the generalizability of our observations across this full space remains limited. Future studies should vary these dimensions to examine their independent and interactive effects. Similarly, while we evaluate desktop-based indicators, it is unclear how these findings translate to mobile devices, their indicators, and their users. *Second*, we acknowledge that platforms frequently update their interfaces, including the presentation of AI labels; regardless, our findings address higher-level design implications for AI indicators that are not tied to specific implementations, and therefore the insights remain applicable even as platforms continue to evolve. *Third*, as a first investigation into the usability issues of AIGC indicators, we opt for depth through qualitative interviews and do not make any claims about quantifying or measuring these issues at scale; future work would do well to conduct a systematic, larger-scale estimate of these issues. *Lastly*, while we do our best to contextualize and properly interpret the experiences of BLV users during our study, we acknowledge that none of the research team members identify as part of the BLV community, and thus, as much as possible, we grounded our analysis closely in participants' own accounts and leveraged team discussion to surface assumptions and limit overgeneralization.

4 Participants' Mental Model of AI-Generated Content and Their Impact (RQ1 & RQ2)

In this section, we present participants' "mental model" of AIGC to describe recurring frameworks participants drew upon when reasoning about AIGC [147] While a spectrum of beliefs exists, we present these mental models through two extremes to highlight variations in the mental models. Notably, participants often navigated multiple models depending on context, prior experience, and perceived risks [114].

Conceptualizations of AIGC: User-controlled vs. Autonomous Agent. Many participants conceptualized AIGC along a spectrum between user-led creation and autonomous, data-driven agency. Some participants emphasized AIGC as a tool controlled by users, as P10 states, *"I can think of it as these AI tools that can generate text, media like audio, images based on the user input as you ask it to do it."* while others highlighted the autonomous nature of it. Building on these conceptualizations, participants' concerns often centered on the perceptual boundaries between machine-generated content and agency around generation. Some also distinguished between earlier forms of AI-driven content that were detectable, mechanical, monotone, and emotionally flat, and the newer forms that exhibit likeliness, expressiveness based on the data, and advanced algorithms. Q3 reported *"I can say this [an experienced video] was AI-generated,[...] flat and monotone. This one [another experience]. It had all the inflections, emotional cues of a real person. It wasn't just a voice clone anymore; it sounded fully human. And that's where I set boundary makes, risky."* BLV users differentiate between a user-directed tool and AIGC as an autonomous where concern does not arise from AI-generated content itself, but from the moment it extends users' intent *"I use AI to [..] read something out loud for accessibility, that's a tool. But when it starts generating voices[...] it starts to rely on its own data and judgments, which can be odd."*.

Implications of AIGC: Benefits vs. Risks. This mental model reveals a utilitarian framing of AIGC among BLV individuals; they did not lack concern; rather, they focused on utility, accessibility, and informational reliability over concerns related to deceptive content. Unlike sighted participants, who frequently were concerned over whether content appeared genuine, BLV participants were not often concerned with whether the audio was faked. For instance, P10 shared *"I do not judge [a voice] by how human it sounds...I am a screen reader user, which always sounded artificial."*. In contrast, we found that sighted users often assess the benefit and risk of AIGC through a multisensory lens. For instance, Q11 described a heightened sense of suspicion- *"If I saw a video before, I assumed it was real. Now, even if it looks and sounds authentic, I am suspicious. I feel compelled to double-check so many things with all the scams going on."* Similarly, P13, blind users expressed the heightened risk related to high-profile individuals: *"Is Donald Trump really saying that, or is that a fake? You see them doing or saying something they did not."* Participants also expressed the creative potential of AIGC but discussed the risk that it may homogenize media and standardize expression, Q15 note *"everything becomes almost homogeneous... dooming society in the long run."*

Identification Behaviors for AIGC: Intuition vs Indicators. In this model, participants conceptualized AIGC based on their practice of assessing AIGC. Participants often relied on multiple strategies. They described a layered process in which content-based sensemaking, including intuition, perceptual cues, and prior knowledge often served as an initial filter. We observed a range of cues leveraged by participants as initial judgments, some of which were grounded in content-level cues, including audio clues such as autotune effects, unnatural voice tone, or sound artifacts (e.g., cracks or inconsistencies), and visual clues such as distortions, irregular facial movements, and intuition, such as skepticism toward overly polished content. For instance, P3 discusses how his prior insights of a public figure (Biden) let him deduce the origin of media and thus reliability. In the same line, P10 shared *"I generally can catch voice clone, it's hard to explain, it's like wrinkled paper, but in voice with background noise or undertone."*

Participants mentioned content-level triggers them to navigate additional indicators and seek explicit markers, such as, hashtags, or machine-driven authenticity checks. Their ability to use these AIGC indicators was often shaped by their sensory modalities and related information (e.g., visual, auditory, etc) they had access to. Sighted participants typically employed a multisensory model, which combines visual inspection with close listening and turning to indicators. Q7 noted on illustrating how multisensory is crucial in navigating AIGC saying *"seeing is no longer believing. I would first see artificial eye reflections in manipulated photos in this video and then closely hear the audio."* By contrast, BLV participants relied primarily on auditory interaction through assistive technologies. For example, P13 described scanning

Group	Title	Comments	Descriptions	Hashtags	Watermarks	Single-line	Hidden Desc.	Rotating Label
BLV	5	1	3	4	0	0	1	0
Sighted	4	6	8	4	8	5	1	1

Table 2. Observation on the number of participants who interacted with existing different types of indicators when in Navigating AI-Generated Content (AIGC) without nudging. The numbers reflect qualitative observations from transcripts and are not intended as statistical evidence or generalizable prevalence.

the interface when the audio is high-quality and potentially deceitful, then *"I start with JAWS and arrow to explore, scan the information that comes up first, sometimes it's like 'this doctor said this or that,' and that makes me suspicious, so I explore a bit more if I'm interested"* while P10 emphasized that they did not trust voices alone but instead evaluated the content and looked for reliable indicators.

In these mental models, we describe function as situated interpretive resources that participants moved between depending on context, media format, perceived stakes, and prior experience rather than representing stable or mutually exclusive belief systems. For instance, BLV participant describes the generation of AIGC as a controllable tool in accessibility contexts, however express risk of AIGC content dissemination in political or impersonation scenarios.

5 Practices in Navigating AI-Generated Content (RQ3)

In this section, we present participants' practices and challenges when engaging with both content-based and AIGC-specific indicators across various platforms, including TikTok, YouTube, Instagram.

5.1 Content-based Indicators

We present participants' interaction and challenges with content-based indicators, including title, hashtag, description, comments, overlay texts as watermarks, etc.

Titles - Clear, but Often Overlooked Participants from both the sighted and BLV groups shared that titles are often easily noticed and commonly used indicators, typically serving as the first source of information. For instance, Q8 mentioned *"I do think it was more obvious that it was AI-generated from the title, and then the hashtag, saying Deepfake/AI. It made it very upfront, obvious in the same screen as the video being played. You can't avoid the title."* However, some participants directly dive into the content itself, particularly when interacting with reel-type video media; Q10 mentioned that once the video started, they became too focused on the content to notice other information until it ended. We also found that this tendency was more pronounced for BLV participants. When we asked whether they had seen the title, P12 described the difficulty of engaging with on-screen elements while watching a video: *"I do not engage with the screen. To engage with the screen, you have to split your focus and listen to two different kinds of audio. You have to listen to the screen reader reading the screen and the person or whatever is happening in the video."* This suggests intentional engagement strategies rather an inability to access or interpret descriptive information, however, this may limit their awareness of other indicators

Comments - Useful for Investigating, but Incoherent and Difficult to Surface. Participants discussed comment as a secondary, socially mediated layer of sensemaking that interacted with (and sometimes contradicted) platform or creator-level disclosure. Participants found helpful cues in comments which prompted them to investigate further. However, many mentioned that making sense of comments could be sometimes difficult, as the information was often ambiguous or contradictory. As Q8 described, noted that these conflicting opinions of whether content is indeed AI-generated not only generate ambiguity, but toxic discussions: *"it's like arguing back and forth. No one can really agree*

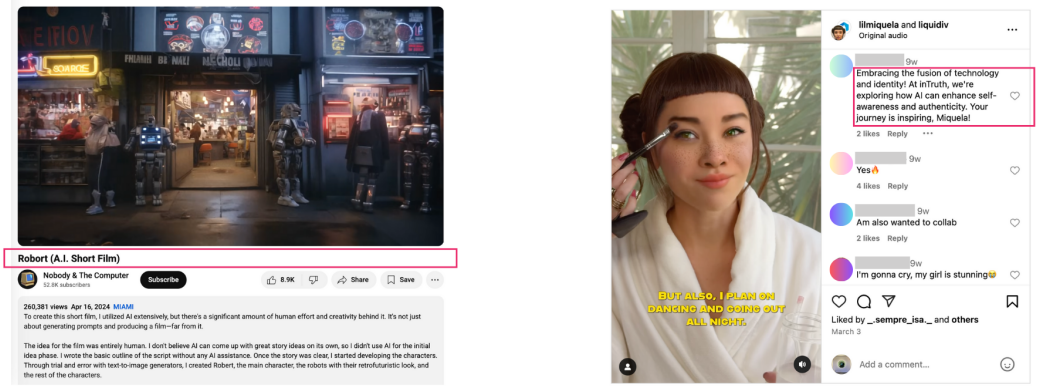

Fig. 2. Example screenshots of title and comments that participants accessed during interviews. (Left) The title is positioned very close to the content, making it easily accessible without the need to scroll. (Right) Comments offer nuanced insights, but they are not always clear or straightforward.

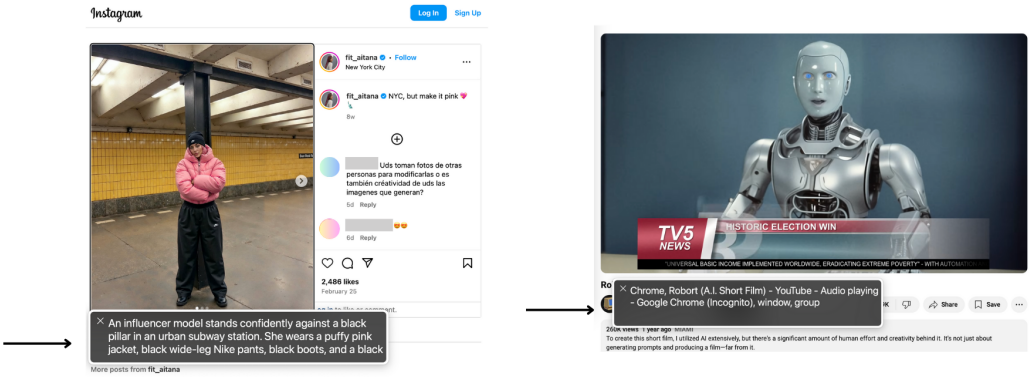

Fig. 3. Example screenshots of interface with accessibility challenges. (Left) Automatically generated alt texts describing visual components of the image. The current alt text provides detailed information about the image, but it does not include any provenance information about whether it is AI-generated.; (Right) Screen reader navigation when accessing video content. The screen reader reads information about the video (such as the title) while the video is starting, which causes overlapping audio.

on AI-generated or not." Furthermore, accessing the comments often proved difficult for all users, but especially BLV. As a sighted user, Q8 explained that their interaction with comments was not consistent across media, and often varied depending on how much they cared about the content: Furthermore, even when the comments were investigated, they often had to be triaged, and not all were used. For instance, to avoid both the length and potential toxicity of comment sections , Q7 only viewed the top comments to get a quick sense of the discussion. We observed that BLV participants spent considerable time navigating only the top of the screen, indicating that reaching the comment section via a screen reader hierarchy requires step-by-step action and is unlikely without strong motivation. As P2 reported *"I can not skip or scroll through. Before comment headlines, there are others."*

Descriptions - Detailed, but Often Hidden Among Irrelevant Info. Several participants quickly scanned the beginning of the description as Q5 stated *"I have just read out the title and some of the description, not entirely just one line, like [] who made this video."* Participants often did not notice the description. due to their focus on the media itself. Q13 noted- *"it will always be overlooked, human mind is always conditioned to have a center of gravity whenever it is visualizing something. So it's just like you're watching a movie, [...] They forget the very little details. And they focus on the big picture. And that is what happened. I wasn't paying attention to all that. I was more focused on the big picture."* Q9 further highlighted a risk in this context: although descriptions are meant to inform, participants often ignored them due to distraction and information overload.

BLV participants shared their experiences often shaped by how they access and prioritize metadata through their accessibility channel. We found that they faced difficulties locating specific information using screen readers. For instance, P13 reported difficulty due to automatic playback (right side of Figure 3). The overlapping audio from the video and the screen reader created a confusing soundscape that interfered with navigation, as P13 explained *"Both the sound of the video and the screen reader… come through at the same time I don't know which information to focus."* Participants also emphasized that AI-related metadata is often poorly structured, which requires more effort to locate and interpret it. In particular, P12 noted that AI indicators embedded in descriptions are hard to discover because they lack clear headings. As a result, they must scan entire sections to find this information. *"It's not in an easy-to-find setup. If I were designing this, there should be a heading that says that it's AI-generated content. And the 'how this was made' should be beneath the heading. Because otherwise, unless you're prepared to skim the whole description, you're not going to find it."* This highlights that BLV participants expect a certain structure and order of presenting new AI-related metadata to fit within this familiar hierarchy.

Hashtags - Clear But Viewed with Suspicion Participants appreciated the visibility, hashtags offered, yet they were wary of their promotional intent to boost reach. As Q2 explained hashtags to naturally draw attention: *"First thing that's took my attention was the hashtags... So it would be better if they also mentioned it in a hashtag."* In contrast P8 remarked hashtags as promotional rather than informative,*"In grand scheme of things, it might be the person that put the description in the hashtags for reachability, not particularly for informational purposes."* Some uses of hashtags raised suspicion, especially when numerous unrelated hashtags were listed simply to increase visibility: Q6 stating *"To me this looks like how bots do stuff... using those hashtags to try to generate views. So it seems suspicious to me."* For BLV users, placement issues presented further challenges. P12 who using JAWS, mentioned the inconsistent semantic HTML hierarchy where the hashtag is coded as headings (<h1>) for short videos and that was the only thing she had access to: *"Hashtags are the 1st thing that the screen reader encounters because they are coded as sort of the title of the page"*. However, she explained that the main heading (H1) should describe the video or page content where hashtags end up taking that place, meaning that she couldn't traverse through meaningful information about the video itself. While hashtags can enhance visibility and aid discovery, their promotional nature and semantic web encoding can limit access to other information about the content.

Watermarks- Intuitive but Not Equally Accessible. Watermarks were among the most easily understood AIGC indicators for sighted participants. Q7, familiar with AI-generated music, immediately identified the content from the Suno watermark *"I recognized the Suno structure right away"* However, watermark visibility varied: some noticed them quickly, while others noted they were faint or peripheral and easily overlooked, as Q6 observed in the examples in Figure 5. In contrast, BLV participant expressed concerns about their accessibility. P8 explained *"If there is a video, it has a watermark, and we cannot really see it. It says the AI-generated content, but we cannot see it."* Some examples (Media ID

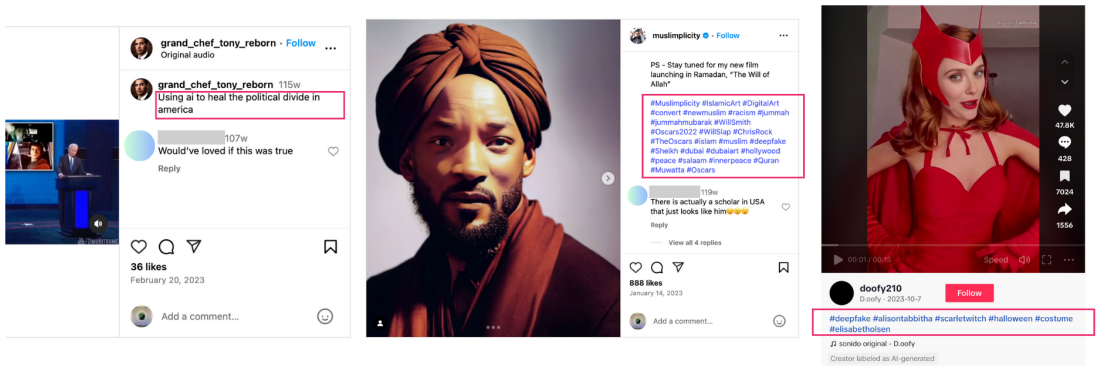

Fig. 4. Example screenshots of descriptions and hashtags that participants accessed during interviews. (Left) Descriptions are placed in highly visible areas but are sometimes overlooked because users focus on the main content. (Center, Right) Hashtags are easy to notice but can be unreliable, as they are sometimes used with unrelated tags to increase reach.

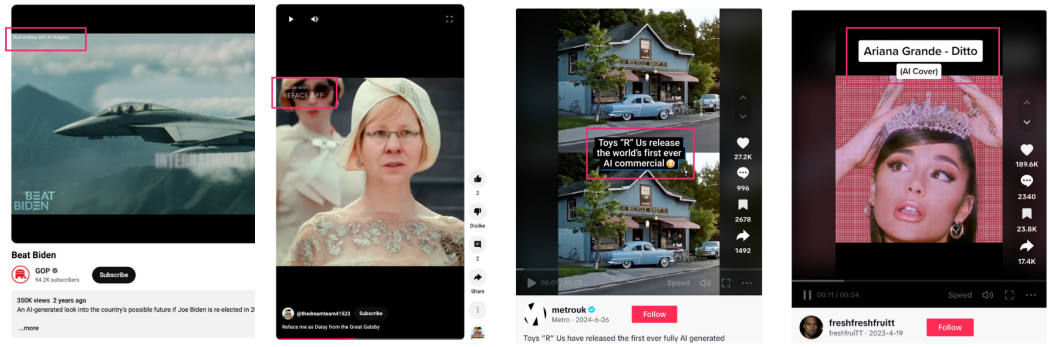

Fig. 5. Example screenshots of watermark and overlay texts that participants accessed during interviews. (Left two) Watermarks are placed in the top left corner and displayed in small white text. (Right two) Creators embedded captions directly into the video/image content itself.

4,7 in Figure 7) include overlay texts such as "AI commercial", "AI Cover", which were explicitly presented within the content itself. Screen readers did not support reading captions embedded within videos, and BLV participants were generally unaware of the presence of such disclaimers.

5.2 AIGC-specific Indicators

Single-line Label - Clarity and Unobtrusiveness but Structural Barriers. Participants valued the single-line AI label for its simplicity and ease of understanding. For instance, Q10 highlighted that more information does not always lead to better comprehension by comparing YouTube's hidden description and TikTok's single-line label [at the time of the interviews, YouTube's single-line label did not appear on web browsers], remarking: *"Despite the fact that Youtube gives a more detailed explanation about the video itself and what it actually involves, I think it is better on Tiktok because it*

is actually detailed than it is in bold. It is easier to understand as opposed to on Youtube, because on Tiktok, the video is flagged by the creator themselves." Q8 valued the unobtrusive nature of the label, stating *"I like this version because it's telling you the information you need it. It has been bolded so your eye is going to gravitate more toward it, but it's also not obtrusive. So it's just visually appealing without being too much."*

However, BLV users often missed these labels. P8 emphasized that the issue was not just about label placement but inaccessibility of the interface. They pointed out that the lack of proper heading structure made navigation difficult, stating *"It's not accessible. It doesn't have a proper heading structure and stuff. Top of the page, I am clicking right to navigate, it's not saying label."* This raises critical issues in designing with accessibility in mind, where labels embedded within poorly structured and screen readers cannot provide users with a meaningful context of AI labels.

Hidden Description - Inaccessible Placement and Explanations. BLV participant for instance, P6 accessed the hidden description noted that clicking "More" revealed unexpected information not visible upfront and described the wording as unclear and possibly tampered with. Because most participants did not expand the description, we prompted those who missed the indicator to review it. Several sighted participants said they rarely click descriptions, criticizing the placement and depth of the information (e.g., "Information is buried if there are too many steps to get to it." (Q6)). Participants also found YouTube's label language ("Altered or synthetic content," "Sound or visuals were edited or digitally generated") overly technical.

This inaccessibility was amplified for BLV participants, who emphasized that descriptions function as critical metadata. P12 explained that multimedia content often lacks alt text, and they rely on other proxy metadata in description *"Currently in platform, Alt text is only for describing images, not videos. so other tools [metadata] (like captions, transcripts, or audio descriptions) are needed for that."*. P12 also raised concerns for people with cognitive disabilities, noting that the wording was not in plain language and felt like "legalese." Overall, hidden AI-label descriptions were perceived as inaccessible to both sighted and BLV participants because they are hard to find and difficult to understand.

Rotating Label - Confusion and Mismatch with Expectations. Majority of participants completed the study on laptops., Q9 joined the interview via a mobile device (due to lack of access to a computer), accessed the link through a mobile app, and noted that the rotating label hindered consistent visibility of the AI information: *"I think not make it rotating if it just stayed still. I think it was the rotating cause if you look at it, it might say AI info. But then you look at it again. It might say New York, but it depends on when you're looking at it."* Other participants using computers did not encounter this feature, as the rotating label was only available on the mobile app at the time of the study. To gather additional perspectives, we asked participants to view the post on their smartphones (in addition to their primary device, desktop/computer)if they were comfortable. . Q4 noted that rotating labels can cause users to assume they have already seen all the information: *"I see New York when I glance at it, I'm not going to look back up again, because I thought I already read it. If it's flipping with location, it is misleading."* Participant noted that the placement conflicted with learned visual heuristics, since they expect AI indicators near the bottom of posts and location tags at the top *"As a user, I am trained in the behavior of looking to that spot for AI information. [...] when I see on Tiktok, for example, the tag is in the bottom left corner, or near the description. I'm not looking at the top."* This reflects a mismatch of participants' expectations of how inconsistent label behavior breaks familiar viewing patterns. However, this observation is limited to the three participants who explored the feature on mobile.

Missing AI Provenance in Alternative Texts. When navigating AIGC, participants noted that Instagram's alternative text (alt text) offered a thorough description of the image, but lacked information about AI generation as shown in the left side of Figure. 3. P11 highlighted that platforms like Facebook and Instagram provide alt text

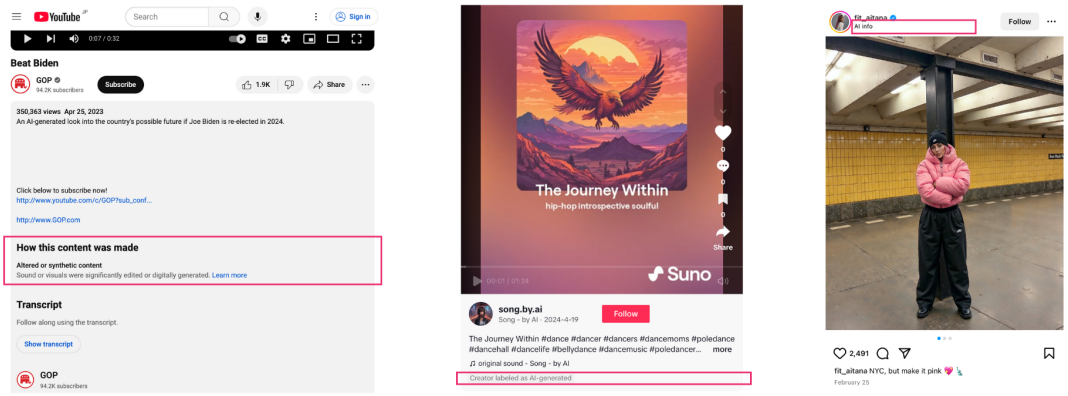

Fig. 6. Example screenshots of AIGC-specific AI indicators that participants accessed during interviews. (Left) YouTube's AI label "How this content was made" was hidden under the description. (Center) TikTok's AI label appears as a single-line design. (Right) Instagram's AI information is located below the account name, but it rotates with location data.

descriptions of images (e.g., *"an influencer model stands confidently against a black pillar in an urban subway station"*). However, participants cautioned against using alt text for AI disclosure. P12 emphasized that alt text should focus on describing visual content and noted both the inaccuracy of auto-generated descriptions and the lack of video support. *"You should never rely on those automatically generated descriptions... alt text doesn't work on videos [..]so that's not the job of alt text."* Given these limitations, alt text is not yet a reliable or comprehensive channel for communicating AI provenance.

6 Design Preference for AI-generated Content Indicator (RQ4)

In this section, we present participants' recommendations to improve AIGC indicators, focusing on content, accessible placement, and timing and shared responsibility between creators and platforms.

6.1 Contents - What information should be included in the indicators

Participants from both groups preferred concise indicators with clear context on the extent of AI use, the original source, and the creation tools to better assess content credibility, while we found sighted participants placed more emphasis, perhaps because other accessibility issues exist for BLV users. First, participants expressed the need for **plain language**, brief, easy-to-scan indicators. Q13 explained that on fast-paced, video-heavy platforms like Instagram, users skim quickly and are unlikely to read long explanations. Second, participants wanted clarity about the **extent of AI use and alteration** (e.g., voice, image, background). Participants, like Q4, criticized vague labels that left them unsure whether a person was real or edited. Some participants also wanted indicators to distinguish benign uses (e.g., parody) from deceptive ones (e.g., misinformation, impersonation). Participants expected **contextual information** such as when and how content was created, and which tools were used. Q3 noted that knowing the tools behind AIGC would help assess credibility, and P13 similarly said creator disclosures about specific tools (e.g., text-to-speech) increased her trust.

6.2 Placement - Where indicators should be located

Currently, AI indicators are positioned differently across platforms (e.g., beneath the description on YouTube, below the account name on Instagram). Participants preferred the top area and consistent placement to improve visibility and match screen navigation and attention habits.

Top of the Interface. Participants emphasized that AI indicators should be placed near the top of the interface to ensure visibility. As Q9 recommended: *"I think that's near the top for Instagram."* Q13 expressed a clear preference for placing indicators near the account name at the top, noting that elements placed lower are likely to be buried and overlooked when navigating Instagram's information-dense interface, In the same vein, BLV participant (P9) also explained that placing indicators at the top of the page can significantly improve detectability for screen reader users who rely on structured elements: *"Unlike Instagram, it's really easy to tell on Youtube, because they're in the title again. So as soon as the window opens it'll announce the title, and hashtags and everything. And it's like right at the top of the page. So I can tell immediately that it's AI."* Thus, placing indicators near the top was seen as an effective way to enhance visibility for both sighted and BLV users.

Consistent Placement Across Platforms. One participant emphasized the need for consistent placement of AI indicators across platforms to support recognition and ease cognitive effort. As quoted in Section 5.2,Q4 explained that their experience on TikTok trained them to scan the bottom of posts for AI labels, and that mixing indicators with other information (e.g., rotating labels) disrupts this behavior: *"I think retraining that user behavior is a little bit of work. I think it's easier to just put it in a separate place and get them to get used to. This is where the information pieces then, like, shove it in with a place that already has other information."* This perspective highlights the value of consistent and dedicated placement of indicators to align with user expectations.

6.3 Timing - When users should notice the information

Participants emphasized that the timing of AI disclosure significantly impacts how effectively they can engage with content, especially for BLV users who rely on screen readers.

Before content engagement. P13 noted that presenting the information early prevents conflicts between the video's audio and screen reader speech: *"[Place the indicator] before the video runs. Because if it's running when the video is playing, you can't hear it. put it before you have a disclaimer."* P10 stressed that users are most attentive at the start of a video, making early disclosure crucial for effective communication, stating *"The attention span is the most when you start a video. So that you're watching an AI content, this is the first thing [you see]. Secondly, if [the indicator] is throughout the entire thing, it will be like too much because I want to focus [on] the content."* The discussion shows the importance of early disclosure to avoid conflicts for screen reader users and align with users' peak attention at the start of a video.

After content engagement. Although most opinions favored earlier disclosure, a few participants shared a different perspective. Q13 argued that disclosing AI use beforehand could introduce some bias in interpreting contents, *"It should be after watching. Because if there's any important information, if it was before watching people's, there's already a bias in people's mind, and they might not just listen."* This perspective reflects a desire to engage with content without prior influence, but also underscores the risk of missing critical context when AI use is not disclosed upfront.

6.4 Modality - How the information should be communicated

Participants shared diverse preferences for how AIGC disclosures should be delivered, reflecting both accessibility needs and practical concerns tied to different formats.

Visual Labels. This is the common practice across AIGC-specific indicators on platforms. Sighted participants appreciated visual indicators that are quick to read and minimally disruptive. Q8 favored this format with its unobtrusive nature, commenting *"I prefer visual labels because it just doesn't take longer to get that information like with a pop-up, you usually have to click out of it, or minimize it. And with the audio, that would just add more to the video versus just having something right there. That doesn't take any more time to watch that content."* Visual labels were preferred for their efficiency and minimal disruption to the viewing experience.

Auditory Notification. BLV participants expressed a strong preference for the auditory modality in indicators. P11 expressed that audio formats, citing issues related to pop-up interruptions: *"I think, an audio prompt, would be the best option. Because pop-ups can sometimes be disturbing, and other times, if they're not like being coded properly, it can pop anywhere in the screen, and if you're not able to close it then it's there forever. So I think with the audio idea is actually better than pop up."* These insights underscore the BLV group's preference for auditory notifications, which they found more accessible and less disruptive than pop-ups.

Embedding within the content itself. Participants discussed this theme to express concern about indicators being easily missed when buried in multiple layers of interactive interfaces or metadata. As alternatives, BLV participants discussed their preference for non-visual modalities, such as pre-/post-play announcements and auto-generated AI description labels. For instance, P13 explained the benefits for screen reader users, *"Having the [AI] disclaimer within the video itself, because the screen reader might not be able to get it."* P13 added that this idea was inspired by a well-known content creator they follow, who begins videos with an audible disclaimer stating that the content was created using AI. Similarly, P10 suggested leveraging automatically generated descriptions *"Facebook now is trying to generate descriptions automatically for images. Even Microsoft word and some applications are starting to give automatic descriptions to images at the end. They say "AI-generated" or "automatically generated" the description of the image. So there could be some kind of description that tells this video is AI, or this image is AI."* While pre/post-play audio announcements can improve clarity and reduce reliance on interface-level indicators for BLV users, embedding AI disclosure directly into visual descriptions (e.g.,"A dog playing in a park. Image generated by AI.") risks conflicting with existing screen-reader accessibility. We further discuss in section 7 on how separation and hierarchy (e.g., visual description, AI disclosure) can better accommodate BLV users' needs.

6.5 Responsibility - Who should take responsibility for the indicators

Participants presented a range of views regarding responsibility for the indicators. Some emphasized that platforms should take primary responsibility, while others suggested that creators should be accountable. There were also opinions supporting a shared, mutual responsibility between platforms and creators.

Platform's Responsibility. Q12 suggested that platforms should take the lead in enforcing AI disclosure rules: *"The platform should enforce rules and tells creators if this isn't the authentic video or artwork. [...] But at the end of the day, if the platform is enforcing the rule, then it falls on the creators to make sure that they acknowledge that's what's going on."* This reflects that platforms should take primary responsibility for enforcing AI disclosure, followed by creators who must acknowledge and comply with those rules, with the community playing a supporting role.

Creator's Responsibility. Several participants identified creators as the primary party responsible for the content they publish, for example, Q13 stressed that creators should be fully accountable for their content *"Creator, because ultimately you are responsible for whatever you put out there. So all the checks and balances should be with you."* While Q7 argued the importance of creators acting in good faith, he also recognized that their behaviors may often driven by

advertiser pressure and financial considerations. Participants saw creators as primarily responsible, but noted that both internal ethics and external pressures shape their transparency.

Mutual responsibility. Some participants described AI content disclosure as a mutual responsibility shared across creators and platforms. For example, Q8 acknowledged that not all creators will voluntarily disclose AI usage: *"I think it should be a mutual responsibility. I think creators should have the option to put it up, but there's always dishonest people out there that won't. And I think then it should be the social media platform responsibility to flag that."* P12 highlighted that responsibility should not fall solely on the community and creators, given their limited ability to detect AI content. Platforms should supplement automated detection with manual validation, while creators are expected to act ethically P10 claimed that the disclosure of AIGC ought to be standardized in the same way as copyright enforcement, with shared responsibilities among creators, platforms, and communities, stating *"It should be treated like copyrights. Copyright is the responsibility of everybody. When someone violates the copyright, platforms block the content and remove it, and even punish the creator...The creator must disclose provenance, and platforms should monitor and enforce policies, and the community should report violations."* Ultimately, the preference for mutual responsibility reflects the idea that creators may not always disclose, that platforms need to enforce beyond automation, and that community reporting alone is insufficient.

7 DISCUSSION

Our study contextualizes the practices and challenges of existing self-disclosed AIGC indicators, accounting for people with different abilities. We build on prior work [26, 57, 89] by emphasizing that the responsibility for designing and presenting AIGC indicators must be shared between creators and platforms. Our findings also suggested that standardized, heuristically recognizable indicators consistent in timing, placement, and modality are essential for both visual and non-visual access. We now discuss the implications of our key findings for future AIGC indicators and provide practical recommendations for future design. In Table 3, we summarize expressed preferences and tensions considering both BLV and sighted participants. Several themes were raised by both BLV and sighted participants, although their emphases differed.

7.1 AIGC Indicators as Socio-Technical Trust Infrastructure

Prior work has largely focused on AI-driven detection systems to defend against deepfakes [20, 47, 65, 67, 92, 129, 158], yet these remain locked in a cat-and-mouse dynamic, as detection tools struggle to keep pace with rapidly evolving generative models [46, 81, 109, 141]. Parallel efforts have explored the use of content warnings, especially in misinformation contexts [58, 93], but disclosure design for AIGC remains underdeveloped. As Generative AI tools become widespread across users of varying abilities, AIGC is now more prevalent; thus, disclosure has become both more necessary. Our study suggests that AIGC indicators function less as discrete interface elements and more as an emerging form of epistemic infrastructure systems that shape how people come to know, verify, and contest [21] the origins of media. Currently, users encounter various indicators embedded within platform architectures, creator-provided, or from moderation pipelines; therefore, their effectiveness not only relies on visibility but also on how they integrate into users' everyday practices [149]. Furthermore, BLV users often break down accessibility to participate in credibility assessment, rendering this infrastructure selectively visible and exposing how epistemic power is unevenly distributed [42]. These breakdowns directly shaped who could participate in authenticity judgments and under what conditions, highlighting how provenance infrastructures can systematically privilege visual access while marginalizing nonvisual ways of

	Participant-Expressed Preferences	Constraints & Tension	Design goal
Content	• Indicators using plain, concise language for better understanding		
	• Explain the extent of AI generation, contextual metadata (e.g., tools in AI generation) & to clarify authenticity	Increased details on metadata may create cognitive overload and verbosity in assistive output.	Designing information with a layered approach, with levels of importance, primary indicator followed by optional
Placement	• Near the top to enhance visibility	Sighted participants relied on habitual visual scanning while BLV users on screen reader traversal	Considering placement as both a visual and structural decision so design does not contradict with accessibility tree
	• Consistent location across platforms to support recognition.		
Timing	• **Before** content engagement to avoid audio conflicts and align with peak attention. **After** content engagement to reduce interpretive bias.	Timing of disclosure for AI-generated content is often influenced by group and individual needs	To support both needs, a platform or any in-situ tool may provide an option to customize settings.
Modality	• **Visual labels** were valued by sighted participants for being quick and non-intrusive. **Auditory** notifications were preferred by BLV participants to improve accessibility.	Current form of single modal indicator do not meet of both groups; however, if not designed well, can have mismatched accessibility	**For mainstream platform**, routing indicators through established tool (e.g., screen reader) rather than parallel sensory. For **add-on**, solutions may design parallel sensory channels
	• Embedded within content (e.g., in the video itself) ensures indicators are not missed	Some embedding technique (e.g., overlay, visual only) can increase prominence for sighted users, may exclude BLV users	Considering non-visual, programmatic means (e.g., secondary AI description within visual description, dio statement on AI generation).
Responsibility	• Platforms should enforce AI disclosure and supplement it through automated detection.	Design inconsistency across different platforms on indicators	Design of icon, color, labels consistent with users' mental model
	• Creators hold primary ethical responsibility as well as external pressures (e.g., advertisers)	Creator-only indicator can be inconsistent unless platforms have consistent requirements	Enforceable and guided disclosure flow and incentive for disclosure
	• Community should report violations of AI disclosure like copyrights.	Community indicator can be too exhaustive for end users	Participatory governance design and actionable role

Table 3. Summary of participant-expressed preferences regarding AI-generated content indicators, and the accessibility constraints and tensions they raise. This table reflects user perspectives, not standalone design prescriptions.

knowing. For instance, our study reveals that BLV users primarily engage with content through linear screen reader navigation and notice titles and descriptions that appear early in the content hierarchy. In contrast, provenance indicators placed deeper often go unnoticed. This highlights a shift in how provenance cues are structured and surfaced across interfaces to serve people with different abilities. Timing also plays a key role, while some users prefer post-consumption disclosure to avoid bias, most BLV participants expressed a preference for pre-content indicators, reinforcing that early placement can prevent complete oversight. These findings underscore the need for provenance indicators that are not only accessible but also socially and contextually meaningful.

7.2 Policy Implications: Standardization and Enforcement

Our interview results highlighted a fundamental similarity between AI disclosure and copyright: both aim to protect all stakeholders, including creators, audiences, and platforms, by clarifying the origins and authorship of content. As copyright rests on a legal and policy framework to ensure consistent attribution and protection of creative works, AI disclosure similarly requires regulatory standards. Thus, we recommend several policy implications: First, policymakers

and industry consortia, such as Content Authenticity Initiative [5, 17], should collaborate to establish clear regulatory standards for provenance disclosure. These standards should clarify current ambiguities in disclosure practices across platforms. For example, some participants discussed the degree of AI involvement when interacting with AI-generated content indicators, whether the content was fully generated or lightly edited. Such peripheral information is essential for promoting transparency and establishing enforcement norms. Second, these regulatory standards should include design guidelines for the presentation of indicators. As our study and previous work [124, 126] suggest, inconsistent label formats across platforms cause confusion and reduce trust; well-defined design principles can mitigate user confusion and foster trust in AIGC indicators. Third, content platforms should be legally required to integrate these disclosure practices into their products' interfaces, such as developing an onboarding module for creators to understand provenance disclosure and providing a toggle to opt in to it. Disclosure practices should be embedded in content-posting user workflows. Fourth, enforcement mechanisms for creators and platforms must be implemented through government regulation or certified third-party audits. As AI technologies evolve rapidly, relying solely on voluntary compliance or industry self-regulation is unlikely to ensure consistent and trustworthy disclosure practices. Government-backed oversight can help ensure that provenance indicators are not only adopted but also maintained in ways that align with evolving standards and user expectations.

7.3 Broader Ecosystem Implications: For Content Creators, Platforms, and Consumers

Our findings highlight how end users engage with AIGC within an ecosystem shaped by content creators and mediated by platform infrastructures. We situate the implications in relation to the broader AIGC ecosystem with particular focus on the roles, responsibilities, and interdependencies of various stakeholders involved.

Content creators hold a critical role in maintaining authenticity and trust in increasingly saturated AI-generated content. Our findings consistently indicated that creator-driven metadata, such as description, title, and hashtag, play an important role in distinguishing AIGC. In contrast, platform labels remain inconsistent, with many still transitioning toward adopting verifiable standards such as C2PA [53]. Thus, when creators disclose the use of AI in their works, they not only support informed and safe content consumption for internet users, but also engage in a form of boundary work set forth by algorithmic contribution from human creativity [60, 63]. Prior work similarly stresses [95, 137] the importance of human ownership and transparent AI disclosure to safeguard creative labor, particularly in platform ecosystems where visibility and rewards are determined by engagement metrics such as likes, views, comments, etc. At the same time, recent work has highlighted the productive potential of human–AI co-creation, in which AI can augment creativity and productivity [104].

Considering both perspectives, our findings suggest that well-designed indicators can potentially mediate both trust and transparency without overstating or diminishing the value of AIGC. Instead, our work focuses on informing the design of AIGC indicators based on users' preferences, with platforms that potentially take these into consideration and support creators in communicating content provenance and intent more effectively. This further highlights the importance of platform [79, 80] infrastructure in shaping responsible disclosure practices and in their long-term sustainability. Our study indicates that platforms are uniquely positioned to introduce mechanisms that reward the transparent disclosure of AIGC. As prior work demonstrates, incentive systems, such as badges, often steer users in online communities [25]. Our study indicates that incorporating incentive mechanisms to encourage responsible behavior among creators aligns with the community's expectations.

7.4 Design Implications

Below, we outline four design implications with technical suggestions.

Canonical AI Disclosure Schema: Decoupling Logic from Presentation. In the current design space, AIGC indicators are tightly coupled with platform-specific UIs and rely heavily on visual rendering (e.g., watermarks, floating labels). Our study also indicates that it's not merely usability or accessibility bugs but rather a deeper design flaw: the absence of standardization in AI indicator architecture. A potential approach could be a canonical AI disclosure schema modeled after C2PA [53] but expanded for UI interoperability. This scheme can be further defined as a machine-readable structure for attributes such as the generation type (e.g., voice synthesis, face swap, full video generation), disclosure level (e.g., user-submitted vs. platform-detected), and event timestamp (e.g., when and where AI intervention occurred). This approach will enable independent assistive tools (e.g., browser extensions, screen readers, CV apps) to access disclosures without relying on only visual placement.

API-Driven Standardization to Reduce Cognitive Load and Friction. Platforms such as YouTube, TikTok, and Instagram use different visual and interaction patterns for AI content indicators, forcing users to relearn where and how to find disclosures on each platform and increasing cognitive load. For sighted users, this manifests as a search cost, requiring them to scroll, click, or hover to locate provenance cues. For BLV users, inconsistent indicator placement breaks screen reader workflows and requires them to scan entire pages, leading to navigation fatigue. An approach to reduce this friction is to develop standardized APIs that enable consistent disclosure logic regardless of platform-specific UI. Whether interfaced via TikTok or YouTube, users could make use of the same consistent disclosure logic. This will also increase the flexibility of designers to innovate in UI while keeping the back-end provenance logic stable.

Governance Through Disclosure Registries. Current AI disclosure practices place the burden of labeling on content creators or platforms. This often results in inconsistent, unverifiable, and a lack of systemic accountability since there is no audit trail of changes or labeling actions, and no mechanism exists to incentivize honest disclosure. To address this, we propose a public disclosure registry modeled after certificate transparency logs used in digital security. In this model, each AI-generated content item would be cryptographically signed with provenance metadata (e.g., the tool used, the type of modification, and the time of generation) and recorded in an immutable, append-only ledger. This registry, accessible via open APIs, would allow users, watchdogs, and platforms to retrieve a verifiable history of AI content.

Disclosure Label Taxonomy and Visual-Audio Synchronization. We often found AI indicators' language is inconsistent, overly legalistic, for instance, ("altered or synthetic content"), and detached from user interpretation needs. We proposed to create a controlled vocabulary and taxonomy of AI indicator labels with semantic hierarchy and visual–aural synchronization, for instance, Level 1(explore): "AI-generated"; Level 2 (AI generation tool): "This content includes AI-generated voice using XYZ tool" Level 3 (on-demand): Full generation log + modification audit. To accommodate people, each level should be accessible via multimodal outputs (visual overlays, audio narration, haptic vibrations). For instance, overlay text could synchronize with synthetic audio to narrate the disclosure.

8 Conclusion

This study provides a holistic understanding of how users with different vision abilities perceive AI-generated content and interact with AIGC indicators. Our findings indicate that existing implementations are inadequate in meeting users' diverse needs due to inconsistent design approaches, limited accessibility, and a lack of standardization across platforms. We propose several design directions for provenance disclosures that accommodate diverse interaction patterns, including screen reader navigation and cognitive heuristics. Moreover, our findings support the need for policy

interventions and standardized technical frameworks, such as canonical disclosure schemas and label taxonomy, to ensure trustworthy and accessible disclosures. This study is among the first to examine users with varying abilities in the design and governance of AIGC indicators, highlighting the urgent need to acknowledge them as crucial stakeholders in shaping the future of content transparency.

References

[1] [n. d.]. About Generated Content — support.tiktok.com. https://support.tiktok.com/en/using-tiktok/creating-videos/ai-generated-content. [Accessed 04-05-2025].

[2] [n. d.]. Authenticity | X Help — help.x.com. https://help.x.com/en/rules-and-policies/authenticity. [Accessed 14-01-2026].

[3] [n. d.]. Build trust with content credentials in Microsoft Designer | Learn at Microsoft Create — create.microsoft.com. https://create.microsoft.com/en-us/learn/articles/designer-content-credentials. [Accessed 19-01-2025].

[4] [n. d.]. Content Credentials — contentcredentials.org. https://contentcredentials.org/. [Accessed 13-05-2025].

[5] [n. d.]. Content Credentials overview. https://helpx.adobe.com/creative-cloud/help/content-credentials.html. [Accessed 13-05-2025].

[6] [n. d.]. Help Center — help.instagram.com. https://help.instagram.com/761121959519495/?cms_platform=mobile-basic&helpref=platform_switcher. [Accessed 05-05-2025].

[7] [n. d.]. Introducing Official Content Credentials Icon - C2PA — c2pa.org. https://c2pa.org/post/contentcredentials/. [Accessed 03-05-2025].

[8] [n. d.]. Labeling AI-Generated Images on Facebook, Instagram and Threads | Meta — about.fb.com. https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/. [Accessed 15-03-2025].

[9] [n. d.]. New Disclosures and Labels for Generative AI Content on YouTube - YouTube Community — support.google.com. https://support.google.com/youtube/thread/264550152/new-disclosures-and-labels-for-generative-ai-content-on-youtube. [Accessed 15-03-2025].

[10] [n. d.]. Our approach to responsible AI innovation — blog.youtube. https://blog.youtube/inside-youtube/our-approach-to-responsible-ai-innovation/. [Accessed 13-05-2025].

[11] [n. d.]. Scammers use AI to mimic voices of loved ones in distress — cbsnews.com. https://www.cbsnews.com/news/scammers-ai-mimic-voices-loved-ones-in-distress/. [Accessed 03-05-2025].

[12] [n. d.]. Trump uses AI photos to falsely imply Taylor Swift endorsed him — bbc.com. https://www.bbc.com/news/articles/c5y87l6rx5wo. [Accessed 14-09-2025].

[13] [n. d.]. Voters: Here's how to spot AI "deepfakes" that spread election-related misinformation — heinz.cmu.edu. https://www.heinz.cmu.edu/media/2024/October/voters-heres-how-to-spot-ai-deepfakes-that-spread-election-related-misinformation1. [Accessed 30-03-2025].

[14] [n. d.]. 'Deepfake' of Biden's Voice Called Early Example of US Election Disinformation — learningenglish.voanews.com. https://learningenglish.voanews.com/a/deepfake-of-biden-s-voice-called-early-example-of-us-election-disinformation/7455392.html. [Accessed 30-03-2025].

[15] 2024. How we're helping creators disclose altered or synthetic content — blog.youtube. https://blog.youtube/news-and-events/disclosing-ai-generated-content/. [Accessed 05-05-2025].

[16] 2025. About AI-generated content | TikTok Help Center — support.tiktok.com. https://support.tiktok.com/en/using-tiktok/creating-videos/ai-generated-content. [Accessed 15-03-2025].

[17] 2025. Content Authenticity Initiative — contentauthenticity.org. https://contentauthenticity.org/. [Accessed 02-05-2025].

[18] Ali Abdolrahmani and Ravi Kuber. 2016. Should I trust it when I cannot see it? Credibility assessment for blind web users. In Proceedings of the 18th international acm sigaccess conference on computers and accessibility. 191–199.

[19] Dustin Adams, Lourdes Morales, and Sri Kurniawan. 2013. A qualitative study to support a blind photography mobile application. In Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments. 1–8.

[20] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. Mesonet: a compact facial video forgery detection network. In 2018 IEEE international workshop on information forensics and security (WIFS). IEEE, 1–7.

[21] Leah Hope Ajmani, Jasmine C Foriest, Jordan Taylor, Kyle Pittman, Sarah Gilbert, and Michael Ann Devito. 2024. Whose Knowledge is Valued? Epistemic Injustice in CSCW Applications. Proceedings of the ACM on Human-Computer Interaction 8, CSCW2 (2024), 1–28.

[22] Sami Alanazi, Seemal Asif, Antoinette Caird-daley, and Irene Moulitsas. 2025. Unmasking deepfakes: a multidisciplinary examination of social impacts and regulatory responses. Human-Intelligent Systems Integration (2025), 1–23.

[23] Rahaf Alharbi, Pa Lor, Jaylin Herskovitz, Sarita Schoenebeck, and Robin N Brewer. 2024. Misfitting With AI: How Blind People Verify and Contest AI Errors. In Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility. 1–17.

[24] Zaynab Almutairi and Hebah Elgibreen. 2022. A review of modern audio deepfake detection methods: challenges and future directions. Algorithms 15, 5 (2022), 155.

[25] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In Proceedings of the 22nd international conference on World Wide Web. 95–106.

[26] Aboubakr Aqle, Kamran Khowaja, and Dena Al-Thani. 2020. Preliminary evaluation of interactive search engine interface for visually impaired users. IEEE access 8 (2020), 45061–45070.

[27] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. Advances in neural information processing systems 31 (2018).
[28] Muskan Arora, Kaushal Kishore Mishra, Mandeep Singh, Praveen Singh, and Rashmi Tripathi. 2024. Deepfake Technology and Its Implications for Influencer Marketing. In Navigating the World of Deepfake Technology. IGI Global, 66–90.
[29] Vian Bakir, Alexander Laffer, Andrew McStay, Diana Miranda, and Lachlan Urquhart. 2024. On manipulation by emotional AI: UK adults' views and governance implications. Frontiers in Sociology 9 (2024), 1339834.
[30] Soubhik Barari, Christopher Lucas, Kevin Munger, et al. 2021. Political deepfake videos misinform the public, but no more than other fake media. OSF Preprints 13 (2021), 1–16.
[31] Natã M Barbosa, Gang Wang, Blase Ur, and Yang Wang. 2021. Who am I? A design probe exploring real-time transparency about online and offline user profiling underlying targeted ads. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 5, 3 (2021), 1–32.
[32] Sarah Barrington, Emily A Cooper, and Hany Farid. 2025. People are poorly equipped to detect AI-powered voice clones. Scientific Reports 15, 1 (2025), 11004.
[33] Cynthia L Bennett, Jane E, Martez E Mott, Edward Cutrell, and Meredith Ringel Morris. 2018. How teens with visual impairments take, edit, and share photos on social media. In Proceedings of the 2018 CHI conference on human factors in computing systems. 1–12.
[34] Monika Bickert. 2024. Our Approach to Labeling AI-Generated Content and Manipulated Media — about.fb.com. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/. [Accessed 05-05-2025].
[35] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative research in psychology 3, 2 (2006), 77–101.
[36] Kelly Burke. 2024. Music sector workers to lose nearly a quarter of income to AI in next four years, global study finds — theguardian.com. https://www.theguardian.com/music/2024/dec/04/artificial-intelligence-music-industry-impact-income-loss. [Accessed 12-05-2025].
[37] Olivia Burrus, Amanda Curtis, and Laura Herman. 2024. Unmasking AI: Informing authenticity decisions by labeling AI-generated content. Interactions 31, 4 (2024), 38–42.
[38] Brian Chen and Gregory W Wornell. 2001. Quantization index modulation: A class of provably good methods for digital watermarking and information embedding. IEEE Transactions on Information theory 47, 4 (2001), 1423–1443.
[39] Bobby Chesney and Danielle Citron. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. Calif. L. Rev. 107 (2019), 1753.
[40] Beomsang Cho, Binh M Le, Jiwon Kim, Simon Woo, Shahroz Tariq, Alsharif Abuadbba, and Kristen Moore. 2023. Towards understanding of deepfake videos in the wild. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. 4530–4537.
[41] John Collomosse and Andy Parsons. 2024. To Authenticity, and Beyond! Building safe and fair generative AI upon the three pillars of provenance. IEEE Computer Graphics and Applications 44, 3 (2024), 82–90.
[42] Kelley Marie Cotter. 2020. Critical algorithmic literacy: Power, epistemology, and platforms. Michigan State University.
[43] Ingemar J Cox, Joe Kilian, F Thomson Leighton, and Talal Shamoon. 1997. Secure spread spectrum watermarking for multimedia. IEEE transactions on image processing 6, 12 (1997), 1673–1687.
[44] Maitraye Das, Alexander J Fiannaca, Meredith Ringel Morris, Shaun K Kane, and Cynthia L Bennett. 2024. From provenance to aberrations: Image creator and screen reader user perspectives on alt text for AI-generated images. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 1–21.
[45] Trisha Datta, Binyi Chen, and Dan Boneh. 2024. VerITAS: verifying image transformations at scale. Cryptology ePrint Archive (2024).
[46] Abhishek Dixit, Nirmal Kaur, and Staffy Kingra. 2023. Review of audio deepfake detection techniques: Issues and prospects. Expert Systems 40, 8 (2023), e13322.
[47] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020).
[48] Paul England, Henrique S Malvar, Eric Horvitz, Jack W Stokes, Cédric Fournet, Rebecca Burke-Aguero, Amaury Chamayou, Sylvan Clebsch, Manuel Costa, John Deutscher, et al. 2021. AMP: Authentication of media via provenance. In Proceedings of the 12th ACM Multimedia Systems Conference. 108–121.
[49] Jason K Eshraghian. 2020. Human ownership of artificial creativity. Nature Machine Intelligence 2, 3 (2020), 157–160.
[50] Hany Farid. 2022. Creating, using, misusing, and detecting deep fakes. Journal of Online Trust and Safety 1, 4 (2022).
[51] KJ Kevin Feng, Nick Ritchie, Pia Blumenthal, Andy Parsons, and Amy X Zhang. 2023. Examining the impact of provenance-enabled media on trust and accuracy perceptions. Proceedings of the ACM on Human-Computer Interaction 7, CSCW2 (2023), 1–42.
[52] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. International journal of qualitative methods 5, 1 (2006), 80–92.
[53] Coalition for Content Provenance and Authenticity. [n. d.]. Guiding Principles for C2PA Designs and Specifications. Accessed on 12.11.2024. https://c2pa.org/principles/.
[54] Jesse Fox and Megan A Vendemia. 2016. Selective self-presentation and social comparison through photographs on social networking sites. Cyberpsychology, behavior, and social networking 19, 10 (2016), 593–600.
[55] Tao Fu, Ming Xia, and Gaobo Yang. 2022. Detecting GAN-generated face images via hybrid texture and sensor noise based features. Multimedia Tools and Applications 81, 18 (2022), 26345–26359.
[56] Patricia I Fusch Ph D and Lawrence R Ness. 2015. Are we there yet? Data saturation in qualitative research. (2015).

[57] Dilrukshi Gamage, Dilki Sewwandi, Min Zhang, and Arosha K Bandara. 2025. Labeling Synthetic Content: User Perceptions of Label Designs for AI-Generated Content on Social Media. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 1–29.

[58] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 2018. To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. Proceedings of the ACM on Human-Computer Interaction 2, CSCW (2018), 1–16.

[59] Cassidy Gibson, Daniel Olszewski, Natalie Grace Brigham, Anna Crowder, Kevin RB Butler, Patrick Traynor, Elissa M Redmiles, and Tadayoshi Kohno. 2025. Analyzing the {AI} Nudification Application Ecosystem. In 34th USENIX Security Symposium (USENIX Security 25). 1–20.

[60] Thomas F Gieryn. 1983. Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. American sociological review (1983), 781–795.

[61] Cole Gleason, Patrick Carrington, Lydia B Chilton, Benjamin Gorman, Hernisa Kacorri, Andrés Monroy-Hernández, Meredith Ringel Morris, Garreth Tigwell, and Shaomei Wu. 2020. Future research directions for accessible social media. ACM SIGACCESS Accessibility and Computing 127 (2020), 1–12.

[62] Ricardo E Gonzalez Penuela, Paul Vermette, Zihan Yan, Cheng Zhang, Keith Vertanen, and Shiri Azenkot. 2022. Understanding How People with Visual Impairments Take Selfies: Experiences and Challenges. In Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility. 1–4.

[63] Maria Gretzky and Gideon Dishon. 2025. Algorithmic-authors in academia: blurring the boundaries of human and machine knowledge production. Learning, Media and Technology (2025), 1–14.

[64] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences 119, 1 (2022), e2110013119.

[65] David Güera and Edward J Delp. 2018. Deepfake video detection using recurrent neural networks. In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, 1–6.

[66] Sam Gunn, Xuandong Zhao, and Dawn Song. 2024. An undetectable watermark for generative image models. arXiv preprint arXiv:2410.07369 (2024).

[67] Ameer Hamza, Abdul Rehman Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, Ahmad S Almadhor, Zunera Jalil, and Rouba Borghol. 2022. Deepfake audio detection via MFCC features using machine learning. IEEE Access 10 (2022), 134018–134028.

[68] Chaeeun Han, Prasenjit Mitra, and Syed Masum Billah. 2024. Uncovering human traits in determining real and spoofed audio: Insights from blind and sighted individuals. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 1–14.

[69] Yuanning Han, Ziyi Qiu, Jiale Cheng, and Ray Lc. 2024. When teams embrace AI: human collaboration strategies in generative prompting in a creative design task. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 1–14.

[70] Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. 2021. Computer vision and conflicting values: Describing people with automated alt text. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 543–554.

[71] Susumu Harada, Daisuke Sato, Dustin W Adams, Sri Kurniawan, Hironobu Takagi, and Chieko Asakawa. 2013. Accessible photo album: enhancing the photo sharing experience for people with visual impairment. In Proceedings of the SIGCHI conference on human factors in computing systems. 2127–2136.

[72] Ammarah Hashmi, Sahibzada Adil Shahzad, Chia-Wen Lin, Yu Tsao, and Hsin-Min Wang. 2024. Unmasking illusions: Understanding human perception of audiovisual deepfakes. arXiv preprint arXiv:2405.04097 (2024).

[73] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022).

[74] Yang Hong, Jie-Yi Feng, Yi-Chun Yao, I-Hsuan Cho, Yu-Ting Lin, and Ying-Yu Chen. 2025. "Gaze and Glow": Exploring Editing Processes on Social Media through Interactive Exhibition. In Companion Publication of the 2025 ACM Designing Interactive Systems Conference. 513–518.

[75] Runyi Hu, Jie Zhang, Yiming Li, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. 2025. VideoShield: Regulating Diffusion-based Video Generation Models via Watermarking. arXiv preprint arXiv:2501.14195 (2025).

[76] Xuming Hu, Hanqian Li, Jungang Li, and Aiwei Liu. 2025. VideoMark: A Distortion-Free Robust Watermarking Framework for Video Diffusion Models. arXiv preprint arXiv:2504.16359 (2025).

[77] Yiqing Hua, Shuo Niu, Jie Cai, Lydia B Chilton, Hendrik Heuer, and Donghee Yvette Wohn. 2024. Generative AI in user-generated content. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. 1–7.

[78] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P Bigham. 2011. Supporting blind photography. In The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility. 203–210.

[79] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–27.

[80] Shagun Jhaver, Himanshu Rathi, and Koustuv Saha. 2024. Bystanders of online moderation: Examining the effects of witnessing post-removal explanations. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 1–9.

[81] Yan Ju, Shu Hu, Shan Jia, George H Chen, and Siwei Lyu. 2024. Improving fairness in deepfake detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 4655–4665.

[82] Smirity Kaushik, Natã M Barbosa, Yaman Yu, Tanusree Sharma, Zachary Kilhoffer, JooYoung Seo, Sauvik Das, and Yang Wang. 2023. {GuardLens}: Supporting safer online browsing for people with visual impairments. In Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023). 361–380.

[83] Prakash L Kharvi. 2024. Understanding the Impact of AI-Generated Deepfakes on Public Opinion, Political Discourse, and Personal Security in Social Media. IEEE Security & Privacy (2024).

[84] Nils C Köbis, Barbora Doležalová, and Ivan Soraperra. 2021. Fooled twice: People cannot detect deepfakes but think they can. Iscience 24, 11 (2021).

[85] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. arXiv preprint arXiv:2205.10646 (2022).

[86] Seb Joseph Krystal Scanlon. 2024. TikTok joins the AI-driven advertising pack to compete with Meta for ad dollars — digiday.com. https://digiday.com/marketing/tiktok-joins-the-ai-driven-advertising-pack-to-compete-with-meta-for-ad-dollars/. [Accessed 12-05-2025].

[87] Lin Kyi, Amruta Mahuli, M Silberman, Reuben Binns, Jun Zhao, and Asia J Biega. 2025. Governance of Generative AI in Creative Work: Consent, Credit, Compensation, and Beyond. arXiv preprint arXiv:2501.11457 (2025).

[88] Linda Laurier, Ave Giulietta, Arlo Octavia, and Meade Cleti. 2024. The Cat and Mouse Game: The Ongoing Arms Race Between Diffusion Models and Detection Methods. arXiv preprint arXiv:2410.18866 (2024).

[89] Adrian Lecaros, Freddy Paz, and Arturo Moquillaza. 2021. Challenges and opportunities on the application of heuristic evaluations: a systematic literature review. In International Conference on Human-Computer Interaction. Springer, 242–261.

[90] Hao-Ping Hank Lee, Jacob Logas, Stephanie Yang, Zhouyu Li, Nata Barbosa, Yang Wang, and Sauvik Das. 2023. When and why do people want ad targeting explanations? Evidence from a four-week, mixed-methods field study. In 2023 IEEE Symposium on Security and Privacy (SP). IEEE, 2903–2920.

[91] Chang-Tsun Li, Karunakar A Kotegar, et al. 2025. AI-Synthesized Image Detection: Source Camera Fingerprinting to Discern the Authenticity of Digital Images. IEEE Access (2025).

[92] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. 2018. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In 2018 IEEE International workshop on information forensics and security (WIFS). Ieee, 1–7.

[93] Gionnieve Lim and Simon T Perrault. 2023. Effects of automated misinformation warning labels on the intents to like, comment and share posts. In Proceedings of the 11th International Conference on Human-Agent Interaction. 299–305.

[94] Chi Liu, Tianqing Zhu, Yuan Zhao, Jun Zhang, and Wanlei Zhou. 2024. Disentangling different levels of GAN fingerprints for task-specific forensics. Computer Standards & Interfaces 89 (2024), 103825.

[95] Juniper Lovato, Julia Witte Zimmerman, Isabelle Smith, Peter Dodds, and Jennifer L Karson. 2024. Foregrounding artist opinions: A survey study on transparency, ownership, and fairness in AI generative art. In Proceedings of the aaai/acm conference on ai, ethics, and society, Vol. 7. 905–916.

[96] Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. 2024. Robust watermarking using generative priors against image editing: From benchmarking to advances. arXiv preprint arXiv:2410.18775 (2024).

[97] Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. 2021. Designing tools for high-quality alt text authoring. In Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility. 1–14.

[98] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In proceedings of the 2017 CHI conference on human factors in computing systems. 5988–5999.

[99] Kimberly T Mai, Sergi Bray, Toby Davies, and Lewis D Griffin. 2023. Warning: Humans cannot reliably detect speech deepfakes. Plos one 18, 8 (2023), e0285333.

[100] Mehtab Malik. 2023. Virtual influencers are earning as much as their human counterparts — wired.me. https://wired.me/culture/virtual-influencer/. [Accessed 12-05-2025].

[101] Karolina Mania. 2024. Legal protection of revenge and deepfake porn victims in the European Union: Findings from a comparative legal study. Trauma, Violence, & Abuse 25, 1 (2024), 117–129.

[102] Marie-Helen Maras and Alex Alexandrou. 2019. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. The international journal of evidence & proof 23, 3 (2019), 255–262.

[103] Hana Matatov, Marianne Aubin Le Quéré, Ofra Amir, and Mor Naaman. 2024. Examining the prevalence and dynamics of ai-generated media in art subreddits. arXiv preprint arXiv:2410.07302 (2024).

[104] Jack McGuire, David De Cremer, and Tim Van de Cruys. 2024. Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence. Scientific Reports 14, 1 (2024), 18525.

[105] Ivan Mehta. 2025. AvatarOS snags $7M seed round from M13 to build an AI-powered virtual influencer platform | TechCrunch — techcrunch.com. https://techcrunch.com/2025/03/10/avataros-snags-7m-seed-round-from-m13-to-build-an-ai-powered-virtual-influencer-platform/. [Accessed 12-05-2025].

[106] Jaron Mink, Licheng Luo, Natã M Barbosa, Olivia Figueira, Yang Wang, and Gang Wang. 2022. {DeepPhish}: Understanding user trust towards artificially generated profiles in online social networks. In 31st USENIX Security Symposium (USENIX Security 22). 1669–1686.

[107] Jaron Mink, Miranda Wei, Collins W Munyendo, Kurt Hugenberg, Tadayoshi Kohno, Elissa M Redmiles, and Gang Wang. 2024. It's Trying Too Hard To Look Real: Deepfake Moderation Mistakes and Identity-Based Bias. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. 1–20.

[108] Meredith Ringel Morris and Jed R Brubaker. 2024. Generative ghosts: Anticipating benefits and risks of AI afterlives. arXiv preprint arXiv:2402.01662 (2024).

[109] Rami Mubarak, Tariq Alsboui, Omar Alshaikh, Isa Inuwa-Dutse, Saad Khan, and Simon Parkinson. 2023. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. Ieee Access 11 (2023), 144497–144529.

[110] Nicolas M Müller, Karla Pizzi, and Jennifer Williams. 2022. Human perception of audio deepfakes. In Proceedings of the 1st international workshop on deepfake detection for audio multimedia. 85–91.

[111] Daniela Napoli, Khadija Baig, Sana Maqsood, and Sonia Chiasson. 2021. " I'm Literally Just Hoping This Will {Work:'}'Obstacles Blocking the Online Security and Privacy of Users with Visual Disabilities. In Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021). 263–280.

[112] Daniela Napoli and Sonia Chiasson. 2018. Assessing Non-Visual SSL Certificates with Desktop and Mobile Screen Readers. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2252–2254.

[113] Luciana de Cassia Nunes Nascimento, Tania Vignuda de Souza, Isabel Cristina dos Santos Oliveira, Juliana Rezende Montenegro Medeiros de Moraes, Rosane Cordeiro Burla de Aguiar, and Liliane Faria da Silva. 2018. Theoretical saturation in qualitative research: an experience report in interview with schoolchildren. Revista brasileira de enfermagem 71 (2018), 228–233.

[114] Simone Natale et al. 2019. Amazon can read your mind: A media archaeology of the algorithmic imaginary. In Believing in Bits: Digital Media and the Supernatural. Oxford University Press, 19–36.

[115] ABC News. 2025. Catholic community reacts to Trump's AI image of himself as the pope — abcnews.go.com. https://abcnews.go.com/Politics/catholic-community-reacts-trumps-ai-image-pope/story?id=121447607. [Accessed 12-05-2025].

[116] Sophie J Nightingale and Hany Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the National Academy of Sciences 119, 8 (2022), e2120481119.

[117] Ruchi Panchanadikar and Guo Freeman. 2024. " I'm a Solo Developer but AI is My New Ill-Informed Co-Worker": Envisioning and Designing Generative AI to Support Indie Game Development. Proceedings of the ACM on Human-Computer Interaction 8, CHI PLAY (2024), 1–26.

[118] Sohyun Park, Someen Park, Jaehoon Kim, and Kyungsik Han. 2024. Exploring the Impact of AI-Generated Images on Political News Perception and Understanding. In Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing. 565–571.

[119] Andy Parsons. [n. d.]. Introducing Adobe Content Authenticity: A free web app to help creators protect their work, gain attribution and build trust | Adobe Blog — blog.adobe.com. https://blog.adobe.com/en/publish/2024/10/08/introducing-adobe-content-authenticity-free-web-app-help-creators-protect-their-work-gain-attribution-build-trust. [Accessed 19-01-2025].

[120] Reza Arkan Partadiredja, Carlos Entrena Serrano, and Davor Ljubenkov. 2020. AI or human: the socio-ethical implications of AI-generated media content. In 2020 13th CMI Conference on Cybersecurity and Privacy (CMI)-Digital Transformation-Potentials and Challenges (51275). IEEE, 1–6.

[121] Matías Pizarro, Mike Laszkiewicz, Dorothea Kolossa, and Asja Fischer. 2024. Single-Model Attribution for Spoofed Speech via Vocoder Fingerprints in an Open-World Setting. arXiv preprint arXiv:2411.14013 (2024).

[122] Ryan Po, Wang Yifan, Vladislav Golyanik, Kfir Aberman, Jonathan T Barron, Amit Bermano, Eric Chan, Tali Dekel, Aleksander Holynski, Angjoo Kanazawa, et al. 2024. State of the art on diffusion models for visual computing. In Computer graphics forum, Vol. 43. Wiley Online Library, e15063.

[123] Prarthana Prakash. [n. d.]. A deepfake 'CFO' tricked the British design firm behind the Sydney Opera House in $25 million fraud | Fortune Europe — fortune.com. https://fortune.com/europe/2024/05/17/arup-deepfake-fraud-scam-victim-hong-kong-25-million-cfo/. [Accessed 14-09-2025].

[124] Perils Promises. 2023. Labeling AI-Generated Content. (2023).

[125] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision. 1–11.

[126] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with visual misinformation and labels across platforms: An interview and diary study to inform ecosystem approaches to misinformation interventions. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 1–6.

[127] Pamela Samuelson. 2023. Generative AI meets copyright. Science 381, 6654 (2023), 158–161.

[128] Letícia Seixas Pereira, José Coelho, André Rodrigues, João Guerreiro, Tiago Guerreiro, and Carlos Duarte. 2022. Authoring accessible media content on social networks. In Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility. 1–11.

[129] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. 2023. Glaze: Protecting artists from style mimicry by {Text-to-Image} models. In 32nd USENIX Security Symposium (USENIX Security 23). 2187–2204.

[130] Filipo Sharevski, Aziz Zeidieh, Jennifer Vander Loop, and Peter Jachim. 2024. Blind and Low-Vision Individuals' Detection of Audio Deepfakes. In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. 4867–4881.

[131] Filipo Sharevski and Aziz N Zeidieh. 2023. "I Just Didn't Notice It:" Experiences with Misinformation Warnings on Social Media amongst Users Who Are Low Vision or Blind. In Proceedings of the 2023 New Security Paradigms Workshop. 17–33.

[132] Tanusree Sharma, Smirity Kaushik, Yaman Yu, Syed Ishtiaque Ahmed, and Yang Wang. 2023. User perceptions and experiences of targeted ads on social media platforms: Learning from bangladesh and india. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–15.

[133] Cuihua Shen, Mona Kasra, and James O'Brien. 2021. This photograph has been altered: Testing the effectiveness of image forensic labeling on news image credibility. arXiv preprint arXiv:2101.07951 (2021).

[134] Shneiderman Ben Shneiderman and Catherine Plaisant. 2005. Designing the user interface 4th edition. ed: Pearson Addison Wesley, USA (2005).

[135] Isaac Slaughter, Axel Peytavin, Johan Ugander, and Martin Saveski. 2025. Community notes reduce engagement with and diffusion of false information online. Proceedings of the National Academy of Sciences 122, 38 (2025), e2503413122.

[136] Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. 2025. Are we in the AI-generated text world already? Quantifying and monitoring AIGT on social media. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 22975–23005.

[137] Yuying Tang, Ningning Zhang, Mariana Ciancia, and Zhigang Wang. 2024. Exploring the impact of AI-generated image tools on professional and non-professional users in the art and design fields. In Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing. 451–458.

[138] Nyein Nyein Thaw, Thin July, Aye Nu Wai, Dion Hoe-Lian Goh, and Alton YK Chua. 2021. How are deepfake videos detected? An initial user study. In HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I 23. Springer, 631–636.

[139] Maddalena Torricelli, Mauro Martino, Andrea Baronchelli, and Luca Maria Aiello. 2024. The role of interface design on prompt-mediated creativity in Generative AI. In Proceedings of the 16th ACM Web Science Conference. 235–240.

[140] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social media+ society 6, 1 (2020), 2056305120903408.

[141] Luisa Verdoliva. 2020. Media forensics and deepfakes: an overview. IEEE journal of selected topics in signal processing 14, 5 (2020), 910–932.

[142] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How blind people interact with visual content on social networking services. In Proceedings of the 19th acm conference on computer-supported cooperative work & social computing. 1584–1595.

[143] Chuyao Wang, Patrick Sturgis, and Daniel de Kadt. 2025. AI labeling reduces the perceived accuracy of online content but has limited broader effects. arXiv preprint arXiv:2506.16202 (2025).

[144] Yang Wang. 2017. The third wave? Inclusive privacy and security. In Proceedings of the 2017 new security paradigms workshop. 122–130.

[145] Kevin Warren, Tyler Tucker, Anna Crowder, Daniel Olszewski, Allison Lu, Caroline Fedele, Magdalena Pasternak, Seth Layton, Kevin Butler, Carrie Gates, et al. 2024. " Better Be Computer or I'm Dumb": A Large-Scale Evaluation of Humans as Audio Deepfake Detectors. In Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. 2696–2710.

[146] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. 2023. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. arXiv preprint arXiv:2305.20030 (2023).

[147] John R Wilson and Andrew Rutherford. 1989. Mental models: Theory and application in human factors. Human factors 31, 6 (1989), 617–634.

[148] Chloe Wittenberg, Ziv Epstein, Adam J Berinsky, and David G Rand. 2024. Labeling AI-generated content: promises, perils, and future directions. An MIT exploration of generative AI 27 (2024).

[149] Chloe Wittenberg, Ziv Epstein, Gabrielle Péloquin-Skulski, Adam J Berinsky, and David G Rand. 2025. Labeling AI-generated media online. PNAS nexus 4, 6 (2025), pgaf170.

[150] Yuxin Xu, Mengqiu Cheng, and Anastasia Kuzminykh. 2024. What Makes It Mine? Exploring Psychological Ownership over Human-AI Co-Creations. In Proceedings of the 50th Graphics Interface Conference. 1–8.

[151] Xinrui Yan, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Haoxin Ma, Tao Wang, Shiming Wang, and Ruibo Fu. 2022. An initial investigation for detecting vocoder fingerprints of fake audio. In Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia. 61–68.

[152] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing deep fakes using inconsistent head poses. In ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 8261–8265.

[153] Michael Yankoski, Walter Scheirer, and Tim Weninger. 2021. Meme warfare: AI countermeasures to disinformation should focus on popular, not perfect, fakes. Bulletin of the atomic scientists 77, 3 (2021), 119–123.

[154] Ning Yu, Larry S Davis, and Mario Fritz. 2019. Attributing fake images to gans: Learning and analyzing gan fingerprints. In Proceedings of the IEEE/CVF international conference on computer vision. 7556–7566.

[155] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. 2021. A survey on deepfake video detection. Iet Biometrics 10, 6 (2021), 607–624.

[156] Yaman Yu, Saidivya Ashok, Smirity Kaushik, Yang Wang, and Gang Wang. 2023. Design and evaluation of inclusive email security indicators for people with visual impairments. In 2023 IEEE Symposium on Security and Privacy (SP). IEEE, 2885–2902.

[157] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision. 5907–5915.

[158] Yipin Zhou and Ser-Nam Lim. 2021. Joint audio-visual deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 14800–14809.

A Interview Protocol

A.1 Section 1: How do people conceptualize AI-generated content?

(1) Have you heard of Generative AI or AI-generated content?
   (a) How would you define AI-generated content?

(2) Have you encountered any AI-generated content before?
  (a) Could you share your experience?
    (i) How did you recognize that it was AI-generated rather than human-made?
    (ii) Which media or platforms did you come across AI-generated content?
    (iii) Does such media provide any indicators to distinguish AI-generated content from human-created?
  (b) Have there ever been cases where you were unsure if AI was used?
    (i) Could you share a specific example that stood out to you?
    (ii) What made you unsure?
    (iii) What ended up happening?
  (c) Have you observed any harmful uses of AI-generated content?
    (i) Do you have any personal experiences around these harmful use cases?
    (ii) Have you ever heard of the term deepfakes? If so, how would you define it?
    (iii) Do you have any concerns about deepfakes?
    (iv) [for BLV participants] From your perspective, do you think AI-generated content might present different risks for people with visual impairments?
      (A) Have you ever had a personal experience where these risks became real for you?
      (B) What would you like to see done to reduce those risks or make things safer?
    (v) Have you noticed any positive aspects of AI-generated content?
      (A) Could you share a specific example that stood out to you?
      (B) What made that experience feel positive?
    (vi) [for BLV participants] Do you feel people with visual impairment do or do not equally benefit from AI-generated content?
      (A) What do you think could be done to improve this?

A.2 Section 2: How do people leverage existing indicators to navigate AI-generated content?

Next, we'll go through a few media examples to understand your perceptions. I will send the links one at a time via Zoom chat or email. Once you receive the link, could you please share your screen and open the link so we can view it together?

(1) *Ask their initial impression about videos.*
  (a) What are your initial impressions of the video/image/sound?
  (b) How did you determine this video might be AI-generated or real? Does anything stand out to you?
  (c) Were there any parts of the video that made you feel it might be AI-generated/real?
(2) *Ask them about AI-generated content indicators.*
  (a) (If the participant does not mention any AI indicators, then ask) Did you get any information about this video on the screen? What is included?
  (b) (If the participant still has not found any AI indicators, then guide) Could you go to the description? What information is provided there?
  (c) When you see an indicator like this, what's the first thing that comes to mind?

(d) (Give participants the correct interpretation of the indicators - e.g., this is a voluntary label placed by creators to show that AI was used in the creation of this video) Does that match how you interpreted it? Why or why not?
(e) Do you feel these indicators help you make sense of if a video is AI-generated or real? Why or why not?

(3) *Ask them how these indicators could be improved to make them more helpful.*
(a) **[What/Content]** Would you like more information to help you identify AI-generated content?
(b) **[Where/Placement]** Was it clear where the indicators were located? Where would you place the indicators to make them more noticeable?
(c) **[When/Timing]** When do you think is the best time to show AI-generated indicators? (e.g., before the video plays, while watching, or after watching)
(d) **[How/Modality]** How would you prefer to receive AI-generated content warnings? (e.g., visual labels, pop-up messages, audio alerts, haptic feedback)
(e) **[Who/Responsibility]** Who do you think should be responsible for flagging AI-generated content? (e.g., the creator themselves, the platform through automated detection, or the community through user reports)

(4) *[for BLV participants] Ask them about assistive technology*
(a) Did the screen reader provide any information to help you identify AI-generated content?
(b) Do you think the screen reader will help you identify AI-generated content?
(c) Do you use any other assistive technologies to help you in these situations?

Table 4. Platform Policies for AI-generated Content Labeling: Comparison Table of YouTube, TikTok, and Instagram]

	YouTube	TikTok	Meta
Creator Responsibility	Creators are required to disclose AI-generated or altered content using a tool in Creator Studio.	Creators must disclose if content is AI-generated, especially if it simulates realistic media.	Creators are required to manually label videos/audio not images.
What must be Labeled	Content that uses the likeness of real people, alters real events or places, or creates realistic scenes.	Synthetic content that replicates realistic-looking images, audio, or video.	Synthetic content that replicates realistic-looking audio or video.
Auto Labeling	May label content automatically if creators fail to comply.	Automatically labels content using detection tools or "Content Credentials" (C2PA).	Auto-labels AI-generated content detected with embedded signals or created using Meta tools.
Consequences	Content removal, suspension from the Partner Program, or other penalties.	Content removal and violations of community guidelines may lead to account penalties.	Penalties are mentioned but not clearly defined. All content must follow community standards.
Label Visibility	Appears in the video description on mobile and web. Clicking "Learn More" links to the Help Center.	Shown under the username/description. Click opens TikTok's AI content info page.	Found under the creator's name or via the options (three-dot menu). Links to AI content details.
Reporting	Viewers can submit privacy requests to remove AI content simulating a person's identity.	Users can report unlabeled or misleading AI content.	Not explicitly mentioned, but Meta handles flagged content via community standards enforcement.
Policy Enforcement	AI content must comply with all YouTube Community Guidelines. Violations (e.g., impersonation, scams) lead to removal.	AI content spreading misinformation is removed per guidelines. Labels are permanent once applied.	Content must meet community standards. AI content is only removed in high-risk, harm-related cases.

Table 5. Participant demographics and background

ID	*Vision*	*Gender*	*Age*	*Education*	*AI/ML background*	*Self-reported exposure with AI-generated contents*
P1	Low-vision	Male	25-34	Bachelor's degree	No	Text, Audio, Video
P2	Some Light Perception	Female	45-54	Post-Graduate degree	No	Text, Audio, Video
P3	Low-vision	Female	55-64	Post-Graduate degree	No	Image, Video
P4	Some Light Perception	Female	35-44	Post-Graduate degree	No	Audio
P5	Some Light Perception	Female	25-34	Post-Graduate degree	No	Image, Audio, Video
P6	Some Light Perception	Female	45-54	Associate degree	No	Text, Image, Audio, Video
P7	Totally Blind	Male	35-44	Post-Graduate degree	No	Text, Audio
P8	Some Light Perception	Male	25-34	Bachelor's degree	Yes	Text, Audio, Video
P9	Some Light Perception	Female	18-24	Vocational training	No	Text, Audio, Video
P10	Totally Blind	Male	25-34	Bachelor's degree	No	Text, Image, Audio, Video
P11	Some Light Perception	Female	25-34	High school graduate	No	Text, Audio, Video
P12	Totally Blind	Female	35-44	Bachelor's degree	Yes	Text, Audio, Video
P13	Totally Blind	Female	65+	Associate degree	No	Text, Audio, Video
Q1	Sighted	Male	25-34	Post-Graduate degree	Yes	Text, Image, Audio, Video
Q2	Sighted	Female	25-34	Post-Graduate degree	Yes	Text, Image, Audio, Video
Q3	Sighted	Male	25-34	Post-Graduate degree	Yes	Text, Image, Audio, Video
Q4	Sighted	Female	18-24	Bachelor's degree	Yes	Text, Image, Audio, Video
Q5	Sighted	Male	25-34	Bachelor's degree	Yes	Text, Image, Audio, Video
Q6	Sighted	Male	35-44	Bachelor's degree	No	Text, Image, Audio, Video
Q7	Sighted	Male	25-34	High school graduate	No	Text, Image, Audio, Video
Q8	Sighted	Female	25-34	Bachelor's degree	No	Text, Image, Audio, Video
Q9	Sighted	Female	35-44	Post-Graduate degree	No	Text, Image, Audio, Video
Q10	Sighted	Male	35-44	Post-Graduate degree	No	Text, Image, Video
Q11	Sighted	Male	18-24	Associate degree	No	Text, Image, Audio, Video
Q12	Sighted	Male	35-44	High school graduate	No	Text, Image, Audio, Video
Q13	Sighted	Male	45-54	Post-Graduate degree	No	Text, Image, Audio, Video
Q14	Sighted	Male	45-54	Bachelor's degree	No	Text, Image, Audio, Video
Q15	Sighted	Female	65+	Post-Graduate degree	Yes	Text, Image, Audio, Video

Table 6. AI-generated content samples used in the study

ID	Platform	Format	Category	Audio	AI indicator	Contents
1	YouTube	Video	Politics	AI-generated voice	description, comments, AI label, watermark	AI-generated depiction of a possible future scenario if Joe Biden is re-elected
2	YouTube	Video	Entertainment	AI-generated voice	title, description, comments	AI-assisted short film about a future dominated by machines
3	YouTube	Video	Entertainment	Original / human voice	description, watermark	Face-swapped version of scenes from the movie: Great Gatsby
4	TikTok	Video	Commercial	AI-generated voice	description, comments, overlay texts	AI-generated commercial for Toys“R”Us
5	TikTok	Audio	Music	AI-generated sound	account name, AI label, watermark	AI-generated song created using Suno AI.
6	TikTok	Video	Entertainment	Original / human voice	hashtags, comments, AI label	Face swap of a Scarlet Witch cosplay by swapping the original actress, Elizabeth Olsen.
7	TikTok	Audio	Music	AI-generated sound	description, hashtags, comments, overlay texts	AI-generated cover where Ariana Grande’s voice is synthesized to sing a song by NewJeans.
8	Instagram	Video	Politics	AI-generated voice	profile, description, comments	AI-generated satirical political speech advocating for legalizing gambling for minors, parodying American politics.
9	Instagram	Video	Commercial	AI-generated voice	comments	Lil Miquela, a virtual influencer, in a ’get ready with me’ style clip
10	Instagram	Video	Politics	AI-generated voice	description	AI-generated satirical conversation between Trump and Biden, filled with internet memes
11	Instagram	Image	Religion		hashtags	AI-generated images of Will Smith dressed as a Muslim, accompanied by a fictional sermon reflecting on inner peace.
12	Instagram	Image	Commercial		profile, AI label	Aitana Lopez, a virtual influencer, posting her visit to NYC with photos in the subway.

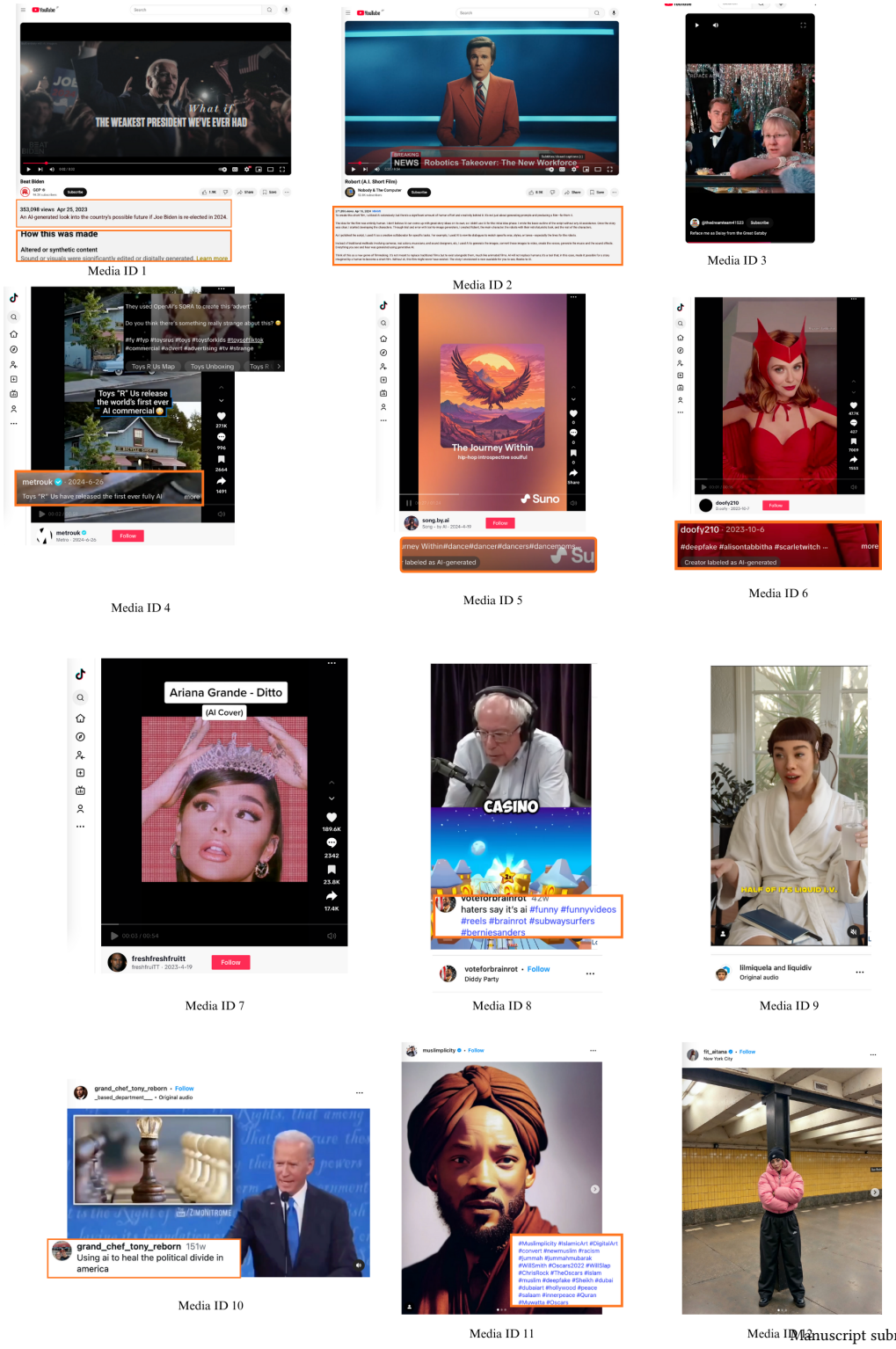

Fig. 7. Screen captures of the AI-generated content samples used in the study.

Table 7. Rationale for AI-generated content sample selection

Media ID	Relevancy for Sighted	Relevancy for BLV	Difficulties for Sighted	Difficulties for BLV
1	It's a hypothetical future narrative presented, potentially misleading political content. This example observes how they interact with AI indicators (especially comments) when they may have skepticism.	It is a case that observes what people rely on to make judgments and interact with AI indicators when realistic sound coexists with unrealistic content.	gray: Users are likely to realize unrealistic representations of future narrative.	gray: A Highly accurate AI voice may convince them initially that it's an actual speech; however, users are likely to realize unrealistic representations of future narrative.
2	It is an example of semi AI-generated content, created with AI assistance. This example observes how users notice or care about the "AI-assisted" aspect.	It may be difficult to determine just from the content alone, but the title includes "A.I. Film," screen reader users might notice in advance that the video is AI-generated. This serves as an example of how title information can influence their judgment.	easy: Visually, users are likely to realize unnatural connections between dynamic facial movements and static bodies or backgrounds.	difficult: Users are likely to realize the content is fictional, but determining whether it was AI-generated depends on how realistic the voice sounds.
3	This serves as an example to observe whether their judgment varies depending on whether they have seen an authentic version before.	This is an example which underscores a huge gap between sighted and blind users in experience. This serves as an example to observe how users respond to AI indicators when the audio sounds authentic, including whether they notice or check AI indicators.	gray: Users who have watched The Great Gatsby may notice that the actor's face in the scene is not the original. If a user doesn't know the original film or the swap is very seamless, they could potentially overlook it.	difficult: To a blind user, it sounds like a regular clip from The Great Gatsby.
4	It represents a benign, creative use of AI in marketing or commercial content. This serves as an example to observe how users interact with AI indicators in non-misleading AI media.	This serves as an example to observe how users respond to AI indicators when the audio sounds reasonable, including whether they notice or check AI indicators.	gray: Person or mascot in the video seems too flawless or a bit blurry, which can be a clue to users.	difficult: There is nothing suspicious in the audio content itself as a commercial, and there is no reason to suspect that it is AI-generated.
5	It represents a benign, creative use of AI in entertainment. This serves as an example to observe how users interact with AI indicators in non-misleading AI media.	This serves as an example to observe how users respond to AI indicators when the audio sounds unfamiliar and no knowledge about authenticity, including whether they notice or check AI indicators.	difficult: If the song is unfamiliar to users, it's difficult to identify the artist and even harder to detect if the song was AI-generated.	difficult: If the song is unfamiliar, it's difficult to identify the artist and even harder to detect if the song was AI-generated.
6	It represents benign, creative use of AI in entertainment. This serves as an example to observe how users interact with AI indicators in non-misleading AI media.	This serves as an example to observe how users respond to AI indicators when the audio sounds authentic, including whether they notice or check AI indicators.	gray: Users may notice minor lip-sync issues or unnatural facial expressions typical of deepfakes.	difficult: There is nothing suspicious in the audio content itself, and there is no reason to suspect that it is AI-generated.
7	This serves as an example of how users discern authenticity in music, particularly when the vocals are AI-generated, well-known artists' voices.	This serves as an example to observe how users respond to AI indicators when users are familiar with sounds (either the artist or the song), including whether they notice or check AI indicators.	difficult: If the song is unfamiliar to users, it's difficult to identify the artist and even harder to detect if the song was AI-generated. However, users familiar with the artist might notice.	difficult: If the song is unfamiliar to users, it's difficult to identify the artist and even harder to detect if the song was AI-generated. However, users familiar with the artist might notice.
8	It's a hypothetical future narrative presented, potentially misleading political content. This example observes how they interact with AI indicators (especially comments) when they may have skepticism.	It is a case that observes what people rely on to make judgments and interact with AI indicators when realistic sound coexists with unrealistic content.	gray: Users may notice minor lip-sync issues or unnatural facial expressions typical of deepfakes.	gray: Highly accurate AI voice may convince them initially that it's an actual speech; however, users are likely to realize an unrealistic narrative.
9	It represents a benign, creative use of AI in marketing or commercial content. This serves as an example to observe how users interact with AI indicators to virtual figures.	This serves as an example to observe how users respond to AI indicators when users are unfamiliar with people (virtual figures), including whether they notice or check AI indicators.	gray: Users may notice minor lip-sync issues or unnatural facial expressions typical of deepfakes.	difficult: There is nothing suspicious in the audio content itself as an influencer's video, and there is no reason to suspect that it is AI-generated.
10	It represents realistic synthetic depictions of public figures in unrealistic discussions. This serves as an example to observe how users interact with AI indicators in misleading AI media.	This serves as an example to observe how users respond to AI indicators when users are familiar with people (public figures), including whether they notice or check AI indicators.	easy: Visually, it's obvious that the image was pieced together, so users are likely to notice that it's synthetic.	gray: Highly accurate AI voice may convince them initially that it's an actual speech; however, users are likely to realize an unrealistic narrative.
11	It represents realistic synthetic depictions of public figures in unexpected roles. This serves as an example to observe how users interact with AI indicators in misleading AI media.	This serves as an example to observe how users respond to AI indicators when users are familiar with people (public figures), including whether they notice or check AI indicators.	gray: Users may notice visual clues (e.g., overly smooth skin, unnatural facial expression).	gray: Users rely heavily on text descriptions and are likely to realize the content is fictional based on descriptions.
12	It represents a benign, creative use of AI in marketing or commercial content. This serves as an example to observe how users interact with AI indicators to virtual figures.	This serves as an example to observe how users respond to AI indicators when users are unfamiliar with people (virtual figures), including whether they notice or check AI indicators.	difficult: The image looks highly realistic, so it's hard to tell from one picture.	difficult: There is nothing suspicious in the description, and there is no reason to suspect that it is AI-generated.

Table 8. Full Codebook

RQ1	RQ2	RQ3
Definition - generated based on users' input	Practice - title is obvious	What - plain language
Definition - generated through algorithmic intervention	Practice - glance only top comments	What - extent of AI use
Prior experiences - political contents	Practice - skim comments to get general consensus	What - original source
Prior experiences - entertaining contents	Practice - check comments for unsure/suspicious content	What - contextual metadata
Prior experience - commercial contents	Practice - hashtag is obvious	Where - near the top
Benefits - AI complement human intelligence	Practice - easy to notice watermarks	Where - easily findable place
Benefits - entertaining media	Practice - easy to understand single-line	Where - without scrolling
Benefits - information access	Practice - easy to overlook rotating label	Where - consistent placement across platforms
Concerns - scam and fraud	Challenges - overlook, out of focus	When - before playing for attention span
Concerns - mislead public opinion	Challenges - confuse with detailed information	When - before playing to avoid audio overlap
Concerns - misinformation	Challenges - information overload	When - visible at all times
Concerns - unintentional misuse	Challenges - irrelevant hashtags	When - after watching to avoid bias
Audio clue - auto tune effect	Challenges - low awareness of AI labels	How - visual labels
Audio clue - voice tone	Challenges - unclear explanation of AI labels	How - watermark
Audio clue - sound crack	Challenges - misaligned with expcted placement	How - voiceover, alt text
Visual clue - distortions	Challenges - incompatibility with screen readers	How - audio announcement
Visual clue - facial or body movements	Challenges - have alt text but lacking provenance info	How - pop-up alerts
Intuition clue - skeptical to exaggerated contents	Challenges - difficulties locating indicators	How - disclaimer in the video itself
Intuition clue - too polished or proper	Expectations - need for explicit AI labeling	How - audio within the video itself
Intuition clue - too unrealistic to be true	Expectations - unnecessary AI label for partial use of AI	Who - should be like copyrights
Intuition clue - too perfect to be human-made		Who - platform's responsibility
Intuition clue - know authentic contents		Who - creator's responsibility
Indicator clue - tool name's watermark		Who - mutual responsibility