AI DISCLOSURE SEMANTICS IN VISUAL DESIGN

Exploring communication, transparency and creator identity issues.

Lucía Medina Galán

Advisor:
Prof. Margherita Pillan

A.Y. 2025 / 2026
Politecnico di Milano
MSc Digital and Interaction Design

AI DISCLOSURE AI
SEMANTICS: AI SEM
VISUAL DESIGN AI V
FOR TRANSPARENCY
AND CREATOR IDENT
CLOSURE AI DISCL
NTICS: AI SEMANTIC
L DESIGN AI VISUAL
ANSPARENCY AI FOR T
REATOR IDENTITY AN
RE AI DISCLOSURE
SEMANTICS: AI
GN VISUAL DESIG

"**Concern for humans** and their future must always be the **chief interest of all technical endeavors**... so that the **creations of our mind are a blessing** and not a curse to humankind."

– *Albert Einstein*

POLITECNICO
MILANO 1863

SCUOLA DEL DESIGN

AI DISCLOSURE SEMANTICS IN VISUAL DESIGN

Exploring communication, transparency, and creator identity issues.

Author: Lucía Medina Galán
Matricola: 241988
Advisor: Prof. Margherita Pillan
Co-advisors: Isabella Ruina, Gabriele Barzilai
A.Y. 2025 / 2026

CONTENTS

LIST OF FIGURES

4. LIST OF TABLES

5. LIST OF GRAPHS

6. LIST OF ANNEXES

ACKNOWLEDGEMENTS

This thesis represents significant academic and personal journey, one that would not have been possible without the support, guidance, and encouragement of many individuals.

I would like to express my deepest appreciation to my advisor, Professor Margherita Pillan, for her guidance, insightful advice, and patience throughout this thesis. I am grateful for the time she invested in my growth and for her ability to provide clarity during the most complex phases of my research.

My sincere thanks also go to Isabella Ruina and Gabriele Barzilai for their generous support and for sharing their knowledge. I also extend my gratitude to Chiara DiLodovico; her professional perspective as a Design Researcher was vital for ths thesis.

I am grateful to Politecnico di Milano for giving me the opportunity to study MSc in Digital and Interaction Design. This program has significantly shaped my perspective as a designer and has provided me with the tools to contribute meaningfully to the discourse on artificial intelligence and transparency.

Finally, I wish to express my heartfelt gratitude to my family: my parents, Eugenio Medina and María Enriqueta Galán, for their support; my aunt, Cecilia Medina; my brother, Carlos Medina; my cousin, Melissa Galán; and my grandmother, Enriqueta Riddle. I would also like to thank my friends Alessandra Forno, Paalini Sathiyaseelan, Francesca Zarate, Gustaf, Kevin López, Santiago Zorrilla, and Alejandro López for their understanding and patience during the intense moments of thesis writing.

ABSTRACT: ENGLISH

Within the social media ecosystem, the traditional concept of authorship is challenged in an era where AI-generated content is becoming increasingly prevalent. Regulatory frameworks such as the AI Act impose transparency through labeling obligations, and in response to these, current disclosure practices on platforms rely on binary categorizations (e.g., "AI-Generated").

This oversimplified approach fails to capture the spectrum of AI-assisted or hybrid creative workflows, generating tensions between legal compliance and the reality of content creators. Existing research focuses primarily on how audiences react to these labels. However, industry findings suggest that content creators are willing to be transparent about how much AI they use in their work, provided that platforms offer accurate and detailed labels. This research explores the criticalities connected to creators' reactions in facing labeling obligations. Through a qualitative study combining think-aloud protocols and interviews with content creators (e.g., photographers, digital artists), this research explores their reactions to specific label wordings (e.g., "AI-Generated","AI-Assisted," "Co-created," "Made with AI") across four content contexts: news, digital art, interior design, and speculative art.

The findings reveal that creative and visual designers struggle with labels and with the fact that these can be interpreted as semiotic signs of authenticity and authorship, producing an impact on the final perception of content. Their concern is that the introduction of labels with ambiguous terminology replaces and nullifies the recognition of human ideation — a non-negotiable core of creative identity.

The study demonstrates that mandatory labels cause confusion and that a more granular approach is needed, one that considers the complexity of hybrid workflows. In conclusion, the effectiveness of disclosure practices must move beyond binary labels toward more specific designs capable of accurately indicating the creation process, clearly indicating the contributions of human intent and AI involvement.

ABSTRACT: ITALIAN

Nell'ecosistema dei social media, il concetto tradizionale di autorialità è messo in discussione in un'era in cui i contenuti generati dall'IA stanno diventando sempre più numerosi. I quadri normativi come l'AI Act impongono la trasparenza attraverso l'imposizione di etichette, e in risposta a questi, le attuali forme di esplicitazione in uso nelle piattaforme si basano su categorizzazioni binarie (ad esempio, "Generato dall'IA").

Questo approccio eccessivamente semplificato non riesce a catturare lo spettro dei flussi di lavoro creativi assistiti dall'IA o da forme di lavoro ibride, generando tensioni tra la conformità legale e la realtà dei creatori di contenuti.

Le ricerche esistenti si concentrano principalmente su come il pubblico reagisce a queste etichette. Tuttavia, i risultati del settore suggeriscono che i creatori di contenuti sono disposti a essere trasparenti su quanta IA utilizzano nel loro lavoro, a condizione che le piattaforme offrano etichette accurate e dettagliate. Questa ricerca esplora le criticità connesse alle reazioni dei creatori nell'affrontare gli obblighi di etichettatura. Attraverso uno studio qualitativo, che combina protocolli think-aloud e interviste con creatori di contenuti (ad esempio, fotografi, artisti digitali), questa ricerca esplora le loro reazioni a specifiche formulazioni di etichette (ad esempio, "AI-Generated", "Assistito dall'IA", "Co-creato", "Realizzato con l'IA") in quattro contesti di contenuto: notizie, arte digitale, interior design e arte speculativa.

I risultati rivelano che i creativi e designer visivi hanno difficoltà nell'utilizzo delle etichette e con il fatto che queste sono interpretabili come segni semiotici di autenticità e autorialità producendo un impatto sulla percezione finale dei contenuti. La loro preoccupazione è l'introduzione di etichette con una terminologia ambigua sostituisca annulli il riconoscimento dell'ideazione umana – un nucleo non negoziabile dell'identità creativa.

Lo studio dimostra che le etichette obbligatorie causano confusione ed è necessario un approccio più granulare che consideri la complessità dei flussi di lavoro ibridi. In conclusione, l'efficacia delle pratiche di divulgazione deve andare oltre le etichette binarie verso design più specifici e capaci di indicare accuratamente il processo di creazione, indicando con chiarezza i contributi dell'intento umano e il coinvolgimento dell'IA.

INTRODUCTION

In recent years, the emergence of Artificial Intelligence has transformed how content is produced and consumed across various the social media ecosystem. IBM defines AI-generated content "as text, image, video or audio created by artificial intelligence models", which aim to assist humans to produce content efficiently and at scale (Mucci, n.d.). Despite these advantages, social media platforms have encountered challenges for users, particularly in distinguishing artificially-generated content (AICG) from user-generated content (UCG). This issue is amplified by AICG's capabilities of high quality and production immediacy, factors that can potentially increase the credibility of misleading information (Corsi et al., 2024). In fact, empirical research demonstrates that a significant percentage of social media users have difficulty identifying AI-generated content(Radivojevic et al., 2024).

An illustrative example of how synthetic media influences users' perception and potentially undermines public trust, is the 2023 "Pope Drip" phenomenon in which the image shows Pope Francis wearing a white puffer jacket. This hyper-realistic meme caused a debate across social media platforms, underscoring the impact of generative AI on social discourse. Specifically, artifical intelligence has evolved, and while harmless at times, can be perceived as authentic (Corsi et al., 2024).

In this regard, emerging regulations, such as the European Union's AI Act (2024) have introduced sets of laws based on the premise that highly manipulated content can undermine users' can compromise users' requirement and entitlement to transparent information; for this reason, content necessitates disclosure via labeling . Social media platforms – Tiktok, Meta and Youtube – are already implementing these measures through AI labeling systems that warn users when the content is synthetically modified.

Nonetheless, Gamage et al. (2025) suggests that while AI labels may improve notions of synthetic presence in media content, their efficacy as a sole solution for maintaining long-term trust and transparency remains limited and inconsistent across platforms. Users' (content creators and audiences) reactions are multifaceted and dependent on the type of content, which contradicts the prevailing 'one-size-fits-all' approach to labeling.

Within social media ecosystems, labels transcend from merely technical affordances and operate as semiotic signs that interact with the complexity of users' mental models. In certain instances,this communicative dissonance regarding labeling practices precipitates unforeseen

repercussions, such as pervasive skepticism, the stigmatization of AI-generated content, or the marginalization of legitimate content.These effects do not only impact platforms, but also advertisers, audiences, and specially, content creators (van Dijck, 2013; Burrus et al., 2024).

In fact, there remains a gap in understanding how content creators themselves perceive and respond to AI labeling systems. Previous studies suggests that autonomou self-labeling is perceived as more authentic and trustworthy than platform-imposed labels (Jung et al., 2025). Nevertheless, the degree to which these labels modulate creator perception remains under-examined, specifically concerning how this practice reframes audience dynamics and the fundamental construction of perceived creative identity.

Burrus et al. (2024) argue that efficacious AI labeling should not only mitigate transparency issues but also reinforce a sense of creative agency. Their study demonstrates that creators would proactively comply with demonstrating the process by which their content was made provided that situational variables, such as audience, their intent and goals are acknowledged. Consequently, Gamage et al. (2025) underscore that wording is a crucial dimension of label design: language such as 'AI-generated ', 'AI-modified', or 'Synthetic' carries different implications regarding the perceived authenticity of the content.

This research explores the criticalities for creators' reactions in dealing with labelling obligations. Through a qualitative study, combining think-aloud protocols and interviews with content creators (e.g., photographers, digital artists), this study explores their reactions to specific linguistic framings (e.g., 'AI-Assisted', 'Co-Created', 'Made with AI') across news and artistic contexts. The aim is to examine how interaction design can support more empathetic and creator-centric labeling practices within social media platforms.

Specifically, the study seeks to address the following research question: In what way does the linguistic framing of AI-disclosure labels for synthetically generated or modified imagery influence creators' perceptions of authorship, their evaluations of representational legitimacy, and their motivations to disclose generative AI usage on digital platforms?

The investigation begins with Part I: Research Background, which establishes the theory regarding social media as a socio-technical ecosystem where current legal mandates under the EU AI Act, and the cognitive mechanisms that drive human perception. The focus then pivots to Part II: State of the Art, where a critical review of existing disclosure patterns identifies the current semantic gaps in how platforms communicate synthetic provenance.

The methodology is detailed in Part III: Study Design, which presents the qualitative approach, by integrating think-aloud protocols and semi-structured interviews, later used to

interview content creators. Consequently, the study reaches its conclusion in Part IV: Findings provides an empirical analysis of the findings, specifically exploring how different semantic examples and linguistic choices modulate creators' sense of authorship and professional identity. Finally, the research culminates in Part V: Design Synthesis, where these insights are translated into communicative messages and label diagrams, such as the 'Spectrum of AI Involvement'. This synthesis facilitates a more nuanced paradigm for transparency, operationalizing creator-centric insights into a semantic understanding regarding AI disclosure.

RESEARCH BACKGROUND

PART I

1 AI LABELS IN THE SOCIAL MEDIA ECOSYSTEM

"Labeling is a commonly proposed strategy for reducing the risks of generative artificial intelligence (AI). This approach involves applying visible content warnings to alert users to the presence of AI-generated media online (e.g., on social media, news sites, or search . . ." (Wittenberg et al., 2024, Abstract, para. 1).

This section explains what a social media platform is composed of and why AI disclosure labels, as part of its structure, are shaped by different stakeholders. Their effectiveness depends on how platforms design interfaces, distribute content, and implement regulatory policies.

1.1 Social Media as Ecosystems of Meaning

Social media platforms can be understood as environments in which human actors, including audiences, creators, advertisers, policymakers, and companies, and non human actors, including interfaces, recommendation systems, moderation tools, and protocol rules, influence how communication is produced, distributed, and interpreted (van Dijck, 2013). From this perspective, platforms are complex structures that operate through technological, political, social, and economic layers that bring users and stakeholders into shared spaces for social and cultural content, where visibility and opportunities for participation occur (van Dijck, 2013; see Figure 1).

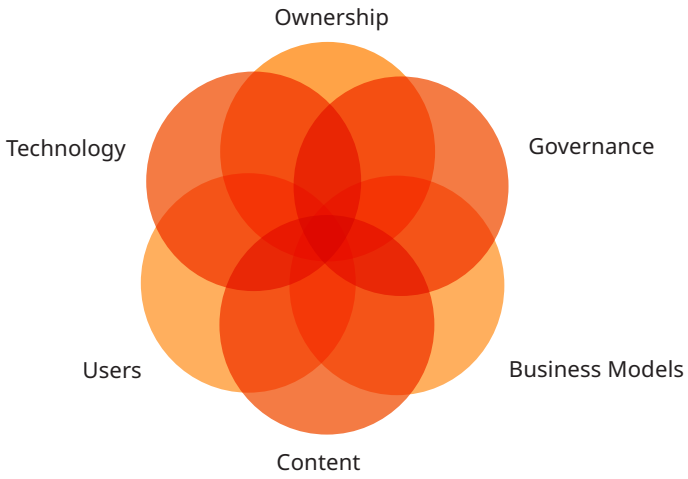


FIGURE 1.1. Diagram dissasembling the social media ecosystem in order to show the interconnection among human and non-human actors (adapted from van Dijck, 2013, p. 28).

Van Dijck (2013) argues that if it is needed to understand an element within a platform, it is necessary to disassemble its components. To understand AI-labels, the author identifies the following structural components through which platforms mediate user experience:

Data

Any type of information such as text, images or sounds. It may also be a nickname, date of birth bio description or gender. It is often provided by the user (p. 30). Additional examples of data include a profile photo uploaded by the user, a Tiktok video or Instagram reel.

Meta-data.

Structured information that explains, describes and locates resources. In other words, the descriptor of the content itself. Examples can be behavioral data such as cookies, keywords of content hashtags or disclosure labels (van Dijck, p. 30). In the context of AI-generated media, metadata can also include embedded information such as AI watermarks or embedded keywords that indicate how content was created. This is different from visible elements such as the Google Studio white diamond logo. For example, Google's SynthID technology embeds invisible watermarks into the pixel data of the image (Bhutani, 2025).

Another example is Content Credentials metadata (powered by the Coalition for Content Provenance and Authenticity or C2PA) that is attached in files generated or edited by AI tools such as Adobe Firefly, DALL-E 3, and Microsoft Designer. Similarly, social media platforms like Instagram and TikTok, employ this feature to automatically label content (Anastasov, 2025, para. 12).

Protocols

Galloway (2024) argues that a core aim of protocols is to control and shape the digital culture, by embedding rules within the platform's interface, they dictate and regulate users' experience and behavior. For Van Dijck (2013, p.31), protocols ar the hidden rules in the background that determine the behaviour of users within the interface such as sharing or joining groups.

Algorithms

Mechanisms that operate behind the interface and are capable of diverse functions and improve software, search engines and social media platforms functionalities. Examples of

algorithm functions are to filter and categore high amounts of data. For instance, Youtube's "most popular videos" category which determines what content is important or relevant to the user (van Dijck, 2013, pp. 49-113).

Algorithms therefore shape how information is presented, and at the same time, they rely on the collective user behavior and on their criteria regarding what counts as high-quality or valuable content (Ciampaglia et al., 2018). An illustrative example of this feedback loop is the case of Twitter's timeline algorithm, which curates a selection of content while simultaneously determining the users choices such as likes, follows and shares (Bucher et al, 2017).

Another functionality of algorithms refers to the optimisation of engagement, which tends to favour sensational, simple and emotionally charged content over accuracy, authenticity or artistic complexity. In other words, this explains how algorithms can contribute to the spread of AI-generated or partially generated content and optimize it for engagement, meaning that the most liked or shared content will be given preference (Liao, 2024).

Interaction cues ('Share' and 'Like' Buttons).

Designed to embed social values. An example is the story of how Facebook embedded the concept of 'connectedness' into their protocol by implementing a 'share' button, and with a similar intention, the feature of the 'Like' button. The 'like' button can also function as a representation of 'support' or sharing information with friends (van Dijck, 2013, pp. 33-54).

Users

Platforms structure information through data, metadata, protocols and algorithms; while users are actors within this social media ecosystem, and play both a passive and an active role by adapting to protocols. Examples of this negotiation are: modifying privacy settings or interacting with algorithmic recommendations that alter content distribution. In this sense, there is a continuous exchange between the users and the elements of the platform (van Dijck, 2013, pp. 12-35). Moreover, users embody two primary roles within the platform ecosystem:

a) Consumers are the receivers of content and are affected by how content is presented and their interactions are defined by protocols and cues.

b) Producers, from amateurs to professionals, they publish, distribute and generate content.

These roles are interchangable; a consumer can be a producer, and viceversa. However, when

a user maintains a constant role as a producer, by becoming an influencer or public personality, they become content creators. These type of users often become more dependent on visibility, engagement metrics, and platform governance decisions (van Dijck, 2013, pp. 14, 38, 74–75).

Within this ecosystem, the concept of sociality is adjusted to the technical aspects of the platform. In this sense, social media platforms re-define the meaning of what is being communicated by structuring how content is presented, connected, and interpreted through the interface (van Dijck, 2013, p. 12).

Therefore, this thesis situates disclosure labels as platform elements that operate both as metadata, data and interface level-protocols that actively shape how content is understood. As data, by describing and signifying the provenance of content; as metadata, by being embedded within the content as a signal of provenance (e.g., Content Credentials); and as protocols, by shaping how information is conveyed to users and and how it should be interpreted. In other words, labels can shape how content is conveyed (Rough & Clift, 2026).

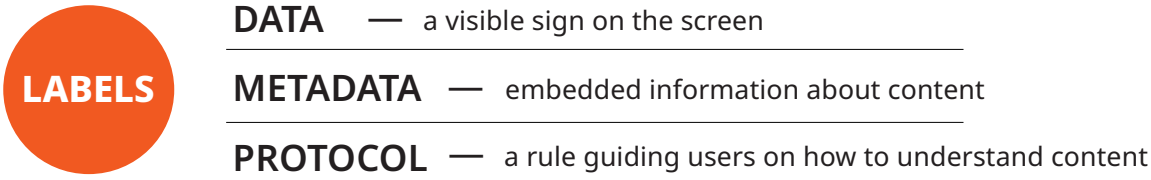


FIGURE 1.1. A label operates as data, metadata, and protocol within the platform ecosystem (visual synthesis based on van Dijck, 2013).

1.2 The Origins of AI-Labelling practices

The act of content labeling on social media is recent and aims to educate users about the provenance of information available within the ecosystem of online platforms. This practice draws inspiration from previous methods in traditional sectors. For instance, the food industry has extensively researched the efficacy of labels to inform consumers about their food choices. Nowadays, policymakers and researchers increasingly orient this approach onto social media platforms, where disclosure of AI-generated content has become an urgent measure to reduce misinformation risks, given that platforms are central sources of public information (Messer et al., 2017; Liao, 2024). In fact, Morrow (2022) describes AI-labels as

tools for disclosure that platforms use with the aim of protecting the human right of legitimate and transparent information. Examples of current regulatory frameworks that enforce AI labelling practices on social media content are the EU AI Act in Europe, and in U.S.A., the U.S. AI governance body.

1.3 What Type of Content needs AI-Labels?

For online platforms, content is a crucial component within the social media ecosystem that connects user agency with platform tecnologies. It can be produced and distributed by users, and it is multimodal, encompassing videos, photos, text, and music (van Dijck, 2013, p. 35). To determine what type of content needs labeling, understanding the umbrella of what constitutes AI-generated content is crucial.

Generative AI content, therefore, can produce content at scale and offer personalization, often outputting sophisticated outcomes. IBM (n.d.) distinguishes three broad modalities:

(a) Text: Generative models understand the rules of grammar and can produce written content.

(b) Images and video: AI can generate, edit and enhance visual content rapidly and at scale,

(c) Sound, and Music: AI systems can simulate natural sounding voices or music that can be blended with other media formats.

Beyond fully synthetic media, AI has also been useful in assisting human creativity by providing tools (e.g., MidJourney, ChatGPT, DALL-e) to refine and edit human-made content and generate ideas. However, current difficulty in determining what is human-made and what is precisely AI-generated destabilizes conventional notions of what can be considered authored. The extent to which generative AI models contribute to a given piece of content lies on a blurry spectrum that ranges from fully automated to lightly modified (French, 2025). The Partnership on AI (PAI), explicitly recognizes the challenge of establishing a clear disclosure threshold within the spectrum. According to a broad consensus of researchers, “a better-contemplated taxonomy for a responsible AI definition is one of the most significant milestones in advances towards HCAI”, without it, there is no universal category that determines to which extent the user can delegate their own responsibility to the system” (Ozmen Garibay et al., 2023, p. 402).

Therefore, AI content can be categorized based on its level of intervention in alignment with current regulatory and platform standards (e.g., EU AI Act, YouTube, Meta):

AI-assisted content (low intervention)

Refers to content using minor enhancements such as color corrections, grammar correction, stylizations, or animations. So far, this type of content, according to the AI Act, does not demand to be labeled/disclosed since the human is the author (e.g., spell-check, grammar correction, noise reduction on a video). In fact, Youtube does not apply any warnings about this content. However, there have been cases where AI-assisted content resulted in being mistakenly labeled as fully synthetic by platforms due to residual Content Credentials metadata (C2PA) – such is the case of Meta, where a slightly AI-edited photograph received an AI Info tag (Anastasov, 2025).

AI-enhanced media / Co-Created (hybrid)

Refers to the middle of the spectrum where there is more debate regarding when and how to label (PAI, 2024). This category applies to content that applied AI-Tools to significantly refine ideas or co-create media:

Filters and auto-enhancements: Complex artistic filters that change the aesthetic of a photograph, or using a language model to generate captions or background audio that influence how the main content is interpreted (Anastasov, 2025).

Co-creation Tools: In fields like music, a human input like voice (core creative idea) while an AI co-creation tool generates the background music. Here, the final piece is a hybrid, where both the human and the machine contribute (Zacharakis et al., 2021). However, the questions regarding who is an author, AI or Human, is still subject to ongoing research (He et al., 2025).

Fully AI-generated (full intervention)

The clearest category for mandatory labeling is fully AI-generated content. The EU's AI Act places specific emphasis on this category due to its synthetic nature, despite the fact that it can be benign (scalable, efficient), since it poses the highest social and ethical risk if transparency is not maintained. In fact, deepfakes are a prime example of synthetic content, enlobing fake videos, audio or images that show events that never happened.

Therefore, disclosure labels on this content type represent a strict measure by platforms and regulatory frameworks due to the fact that social media content can appear indistinguishable from legitimate sources, posing a risk to consumers, individual identities (celebrities, politicians) or nations who are susceptible to misleading narratives (Kharvi, 2024). A notorious

example of deepfakes with malicious intent is the fake video that circulated around social media of President Zelenskyy asking his people to surrender to Russia (e.g., the 2022 deep-faked video of President Zelenskyy) (Liao, 2024).

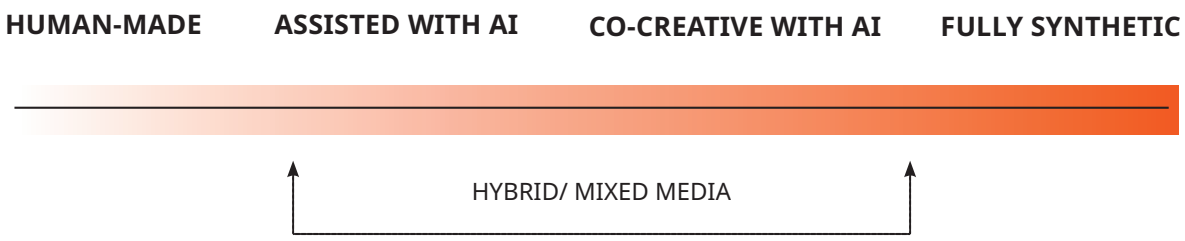


FIGURE 1.3. Blurry Spectrum of content creation (adapted from The spectrum of AI-generated content by the Australian government department, Science and Resources, 2025).

However, determining when content should be labeled, whether minor AI involvement or hybrid involvement, represents a challenge. As the spectrum shows (view Figure 1.1.3), content creators may use AI to different extents, and these distinctions are not always clear among platforms or audiences and creates practical difficulties, which will be defined in the following section. If AI content is an expression of media visualiation, AI-labels are supposed to translate and communicate the provenance of this data.

1.4 The implications of AI-labels for Content Creators

The previous sections outlined the technical definitions of AI-generated content. However, the practical application of AI labels on social media plaftorms introduces social risks for content creators. Within the social media ecosystem described by van Dijck (2013), creators are producers of content who operate within a "reputational economy," where their success depends on producing content, maintaining trust, authenticity, and engagement metrics (likes, shares, visibility).

The Challenge of Standardization vs. Nuance.

Despite the growing agreement that some AI-generated content should be labeled, there is still uncertainty among social media platforms regarding how to implement labeling efficiently while balancing impacts across different stakeholders (Wittenberg et al., 2025). Social media platforms collapse diverse forms of AI involvement into a single category and limit opportunities for creative workflows (Burrus et al., 2024). He et al. (2025) found that people

do not take a one-size-fits-all approach to attributing AI for different contributions. Instead, audiences assigned different types of credit depending on the type of AI-Human contribution, the amount of material produced by AI, and whether the AI acted proactively.

For this reason, a standardized and nuanced taxonomy for AI-generated content is needed, specifically regarding the spectrum of AI involvement: AI-assisted (low AI intervention), Co-partnership with AI (hybrid intervention), and fully synthetic media (e.g., deepfakes). Additionally, current definitions of authorship face challenges in an era where AI involvement is becoming a tool to support human creativity (Burrus et al., 2024; He et al., 2025; Ozmen Garibay et al., 2023; European Parliament, 2025). Because creators act based on an "imagined audience," a single label can be interpreted in ways the creator did not expect, increasing reputational risk and making disclosure feel unsafe or strategically unclear (Marwick & Boyd, 2011).

The Effects of One-Size-Fits-All Labels

Audience interpretations are contextual and depend on the content type. In a study by Burrus et al. (2024), participants interpreted AI use negatively in news content, while finding it more acceptable in other content contexts (influencer marketing). Additionally, the authors found that the perceived level of AI contribution had a direct impact on how participants judged the content's authenticity; when content was perceived as only slightly edited with AI, authenticity judgments were less affected.

This explains why a one-size-fits-all approach to labeling practices is limiting. Consequently, content creators are then left with ambiguous expectations about whether their own content should be labelled as 'AI-generated' or not (Burrus et al., 2024). Furthermore, perceived AI use can negatively affect perceptions of the creator without improving content judgments, disclosure labels risk creating reputational costs while offering limited interpretive benefit (Rae, 2024).

Overall, these challenges and side-effects demonstrate that labelling practices face structural challenges. Therefore, a taxonomic approach focused on granular attribution about AI contribution to hybrid or AI-assisted works; not only for written content, but for images and videos too is needed (He et al., 2025).

1.5 Chapter Summary

This chapter establishes a foundation for understanding AI-labeling within the broader ecosystem of social media platforms and examines the emergence of AI-labeling as a regulatory response to the growth of synthetic media and its unintended challenges such as: The difficulty in identifying ambiguous thresholds between AI-assisted, AI-enhanced, and fully AI-generated content; inconsistencies and errors in automated detection systems and frictions introduced into creators' workflows and identity. These tensions highlight the fact that labeling is not a purely technical mechanism, but rather one that is socio-technical in nature and is affected by platform governance, creative authorship, user perception, and regulations.

2 ETHICAL AND LEGAL FRAMEWORKS

"In the case of social media platforms, it is unavoidable to integrate economic and legal structures as formative factors from the very onset" (van Dijck, 2013, p.27).

This chapter examines the complexity of labeling as a disclosure mechanism showing how it intersects with ethics, regulatory frameworks, platform implementation and conceptual confusion around authorship.

2.1 Foundational Ethical Principles Of AI Disclosure

Nowadays, deepfakes can be easily generated without cost, and this accessibility gives malicious actors the ability to spread convincing synthetic events, impersonate individuals or manipulate public opinion at speed and scale (Boediman, 2025) .

For this reason, both Boediman (2025) proposes a focus on ethical, legal and technological aspects of social media, to show how media credibility is composed of three main ethical pillars: Accuracy, objectivity and transparency.

When these pillars are compromised, individuals are not able to analyze information critically from online sources. Therefore, Boediman (2025) proposes specific dimensions that could ensure that the information conveyed to the public remains reliable:

» **Accuracy and Facts:** emphasizes the importance of presenting verified information,

ensuring that it is authentic.

- **Facts precision:** is about monitoring details and context.
- **Journalistic Ethics:** suggests that transparency, responsibility and avoidance of conflicts of interests should be implemented in the process of presenting sensitive information.
- **Transparency:** suggests that it should be of high importance to show the audiences the contents' production process.
- **Context and Analysis:** this dimension is about the contextual analysis of media to enhance audiences' understanding when information is misleading.

These dimensions offer a view on how social media platforms could preserve the credibility of their content at a moment in time where synthetic or modified information thrive. The 'Transparency' dimension supports the rationale for implementing disclosure labels, emphasizing the importance of showcasing content's production process to the audience. Similarly, Corsi et al. (2024), reinforced this by arguing how deepfakes posed a threat to information ecosystems, stating that policy makers should allow the public to witness the content's production process to prevent exposure to media that is misleading and to avoid the decrease in public trust towards the reliability of visual information.

Similarly, Aler Tubella et al.(2019) assert that ensuring responsible practices towards the usage of AI is not merely about developing systems with trustable outcomes, but about embedding human values into the initial stages of Large Language Models (LLMs)'s development. The authors argue that prevailing approaches risk biases in the information injected to algorithms. A framework for responsible AI, called the "Glass Box" model, is proposed to contrast the metaphor of AI's "black box", where system's choices are made explicit and ethical boundaries are made visible and traceable from input to output stages (Aler Tubella et al., 2019).

As figure 2.1 shows, the Glass Box model is proposed as a governance framework where the interpretation stage translates human values, such as transparency and fairness, into the platform ecosystem. These values are embedded through a top-down process into the design requirements of the AI system. Then, at the observation stage, the AI system is continuously monitored and traced (Aler Tubella et al., 2019).

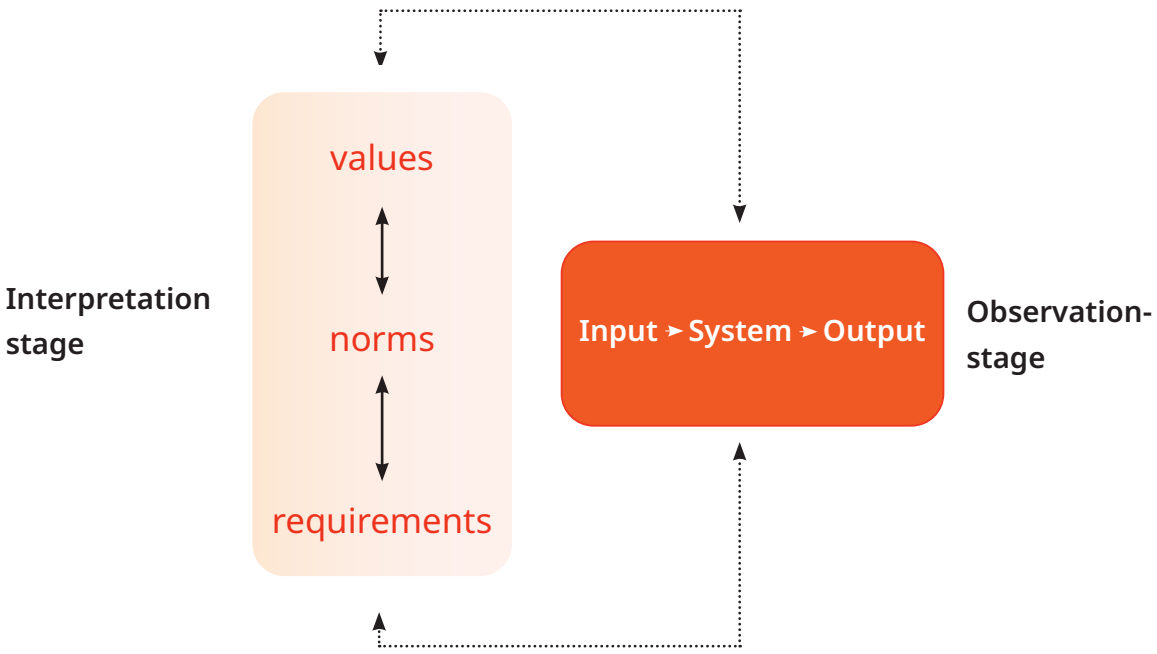


FIGURE 2.1. The Glass Box model of Responsible AI. Adapted from Aler Tubella et al. (2019).

Moreover, the focus on responsible development and use of AI helps safeguard human-centric decision-making in cultural contexts (French, 2025). For this reason, ethical principles are evolving to offer normative direction towards AI. In fact, regulatory frameworks operalize these principles into disclosure mechanisms, such as content labeling on synthetic media, to ensure transparency, accountability and accuracy. Labels aim to offer audiences context about the provenance of content (transparency and accuracy), while simultaneously protecting content creators freedom of expression (accountability) (The Oversight Board, 2025).

2.2 Transparency as governance

The concept of transparency is linked to an ideal, since it is not simply a "precise end state" in which everything is disclosed, but a form of control over a specific stage of a process. Therefore, transparency can be a tool for governance (Annany & Crawford, 2018).

In the era of AI, what kind of transparency is being demanded?

Tubella et al. (2019) introduced the concept of the "White Box" as a proposal for revealing the process of AI generation, by translating human values, such as transparency, into the platform's ecosystem by showcasing the full generation process. Similarly, Boediman (2025)

emphasizes on transparency as a foundational pillar for the ethical, legal aspects of social media platforms in order to disclose the full process behind deepfakes.

However, other studies reveal that disclosing the content's full production process to the audiences, can be counterproducent. While the act of disclosure might promote positive attitudes towards a system, it does not increase expectations of trustworthiness, and consequently might erode them. When implementing transparency and algorithmic awareness on interface design, previous research about user's perception needs to be assessed, since full transparency can be harmful and create false binaries or radical perceptions (Kizilec, 2016; Gamage et al., 2025).

Transparency is an inadequate tool for governing algorithmic systems because it treats AI as static, when in truth, it is a socio-technical assemblage between its developers, the companies that own it, the rules they follow, and the data they choose to show to the audience (van Dijck, 2013, p.27; Ananny and Crawford, 2018).

Consequently, the "disclosure devices" (Hansen and Flyverbom, 2015) used to provide large-scale transparency, such as labels, are not objective claims of data, but structures that decide how information is made visible and how it can be observed often selecting what to reveal.

2.2.1 The EU AI Act

The EU AI Act requires transparent disclosure mechanisms to promote awareness of the presence of AI, such as deepfakes (Artificial Intelligence Act, 2025). For instance, this regulatory framework introduces a set of rules that should be appleied depend on the level of risk AI systems pose to human rights. There are four types of risks (Artificial Intelligence Act, 2025):

1. **Unacceptable risk.** AI systems that fall in this category are entirely banned if their intention is to manipulate through subliminal techniques, exploit vulnerabilities, social scoring, untargeted scraping for facial databases or biometric categorisation (European Union, 2024, Art. 5(1)(a–g)).
2. **High risk.** AI systems that fall into this category can represent a threat to the fundamental rights – safety or health. They include AI used in regulated technologies (medical devices and vehicles), in education (grading), law enforcement, migration and administration of justice (European Union, 2024, Annex III).
3. **Limited Risk.** This category is about transparency obligations in which developers must

notify users about the presence of AI (deepfakes, chatbots). Examples are systems for emotion recognition, systems that generate or manipulate content that could mislead users if not disclosed (European Union, 2024, Art. 50).

4. **Minimal Risk.** This category applies to unregulated applications – such as spam filters and game AIs – that are allowed without obligations. They represent the majority of AI applications (European Union, 2024).

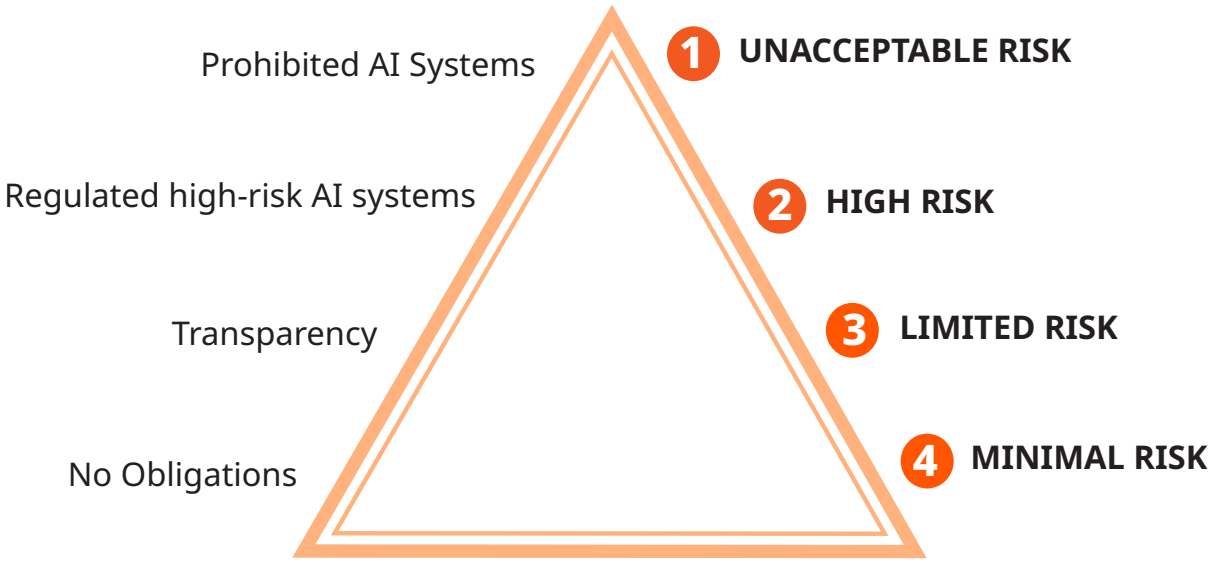


Figure 2.2.1. The AI Act defines four levels of risk for AI systems. Adapted from Regulatory framework for AI, European Commission (2026). Retrieved from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

From the perspective of AI-disclosure labels, Article 50 is of importance. It requires that users: (a) are informed when they interact with an AI system rather than a human, (b) receive disclosure disclosure labels when content has been artificially generated, and (c) are notified when emotion recognition or biometric categorisation systems are being applied to them (European Union, 2024, Art. 50).

Deepfakes are an exemplary use case for this matter. Article 50 (4) of the Act states that :

"Deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake, shall disclose that the content has been artificially generated or manipulated" (European Union, 2024, Art. 50(4), para. 4).

Social media platforms must disclose when content is synthetic, to maintain transparency

and trust. This clause states that deepfakes – or other AI-generated content – must be clearly labeled. However, in cases where the content is artistic, humorous or creative, disclosure can be adapted to avoid interfering with the "enjoyment of the work" (European Union, 2024, Art. 50(4)).

This regulatory framework aligns with Boediman (2025)'s argument that value of transparency should be implemented revealing the production process behing Generative Media. At the same time, the implications of disclosure depend on the type of content. The EU AI Act acknowledges this by allowing adapted disclosure in artistic, humorous or creative contexts, yet does not propose a detailed taxonomy of degrees of AI involvement in creative work. Instead, it emphasizes that fully synthetic or misleading content must be disclosed, and remains less specific about hybrid or AI-assisted content where human authorship is involved.

2.3 Industry Standards and Voluntary Frameworks

Outside governmental regulations, industry coalitions are developing labeling standards. The Partnership on AI (PAI) elaborated a framework to help content creators. Firstly, three categories of stakeholders are defined: builders of technology and infrastructure for synthetic media, creators and distributors/publishers of synthetic media.

The framework is grounded in concepts of transparency, consent and disclosure. It recognises techniques – such as the representation of people, inventing personas, simulating events that never occurred, inserting or removing artefacts from authentic media – can be used responsibly or cause harm. For this reason, PAI emphasises on collaborative research, media literacy initiatives and shared best practices to address the "gray areas" where the impact of synthetic media is difficult to predict.

The framework divides disclosure mechanisms in two categories. Direct disclosure refers to visual cues such as content labels or disclaimers. On the other hand, indirect disclosure, referring to provenance metadata embedded in the file or content itself, such as C2PA, watermarks or pixel-level signals. In fact, PAI encourages tool builders to integrate indirect disclosure by default, and to publish clear policies.

For creators, the framework recommends self-disclosure about the use of AI. This involves embedding provenance tools provided by platforms, and being explicit about ethical boundaries, except in justified artistic or satirical contexts. For platforms, the framework proposes applying labels that are contextual and accurate to the users.

A practical example of implementation of the framework is the case of Tiktok's adoption of AI-labeling policies as part of PAI's case study programme. TikTok uses direct automated disclosure labels (identified as 'AI-generated' or 'AI-edited') with self disclosure features that allow users to mark their own content as synthetic. This is an example of how voluntary standards, such as PAI, are adopted and implemented by platforms. Labels are, therefore, part of a governance strategy that connects creator responsibility, visible interface transparency and provenance metadata.

2.4 Core tensions: Authorship, Copyright, and Creative Integrity

Media production tends to be collective, involving publishers, technicians and infrastructures that affect the final outcome; therefore, authorship is rarely singular and stable. In this thesis, authorship is defined as a social and legal recognized status of being responsible for the creative work. With the rise of generative AI in social media platforms, questions regarding who should be considered the real author (or responsible agent of creative output) highlight a core tension regarding the notion of authorship.

What is the definition of authorship in the era of generative AI?

Foucault (1979) defined authorship as a socio-legal function of discourse that attributes the work to an author in order to enable appropiation, accountability and interpretive control; however, this notion is contructed and fluctuates across time and domains. Consequently,- French (2025) argued that in the AI generative era, the concept of human creativity becomes ambiguous concerning the extent of where it begins or ends.

Bomba and De Angeli (2025) elucidate, through studies on generative AI art, that agency and authorship are relational and distributed across artists, data, and algorithms. The authors frame AI art as produced through interactions between human intention and the machine, complicating the notion of who is the author.

Is AI generated content copyrightable?

Legally, AI-generated content is not considered copyrightable since it lacks originality and is currently based on existing data and algorithms. As French (2025) states, in both US and EU, synthetic content is not copyrightable and is owned by the public. Therefore, for legal purposes, when a content is made by a human and an AI collaborator, the human is considered the author. Furthermore, the UK government, under the UK's Copyright, Designs and Patents

Act, only recognizes fully AI-generated work as copyrightable, but undermines the nuanced spectrum of AI intervention (French, 2025).

While copyright defines who has rights and accountability, platform AI disclosure practices affect how audiences define who deserves recognition. It is therefore useful to distinguish ownership, authorship, and attribution: He et al. (2025) define attribution as the visible recognition given to contributions of a work (in social media, of a certain content), and note that it differs from ownership both legally (rights and accountability) and psychologically (a sense of possession). Moreover, ownership is independent of attribution, and authorship does not guarantee visible attribution. This distinction is relevant for AI-mediated creation, because disclosure labels operate primarily at the level of attribution signals (He et al., 2025).

Figure 2.4. Distinguishing legal ownership, practical authorship, and social attribution; AI disclosure labels primarily affect attribution (He et al., 2025).

In practice, AI tools facilitate the production of media at different stages by making creation more accessible. An example is the case about the GAN (Generative Adversarial Networks') artist – Barrat, who self-describes as a curator/artist. He explained that the final output of a generated image is an inaccurate representation of the input data. This case highlights how GAN's reflect biases in input data, which can lead to results that deviate from the initial original intent (Katja de Vries, 2020).

Therefore, the concept of authorship becomes ambiguous since creative decisions are distributed between human and machine creators. This raises tensions regarding who should receive recognition, who is responsible and what it means to claim a work as 'own' in AI-mediated creative practices. An example of this ambiguity is the case of the artwork called *Théatre D'opéra Spacial* that won a prize during an art competition, and when the authored claimed the legal rights, the USA Regisration Office denied it (Citterio, 2024).

FIGURE 2.4 Jason M. Allen, Théâtre d'Opéra Spatial (2022), digital image (AI-generated/AI-assisted). Source: Wikimedia Commons. Public domain (per Wikimedia Commons file license).

He et al. (2025) examined that different levels of AI-attribution led to different credit claims, and argued that policies and frameworks need to re-examine their structures, specifically regarding co-created content. Furthermore, Xu et al. (2024) suggest that policies should consider AI involvement as a spectrum, and therefore, create laws accordingly. Burrus et al. (2024) further elaborates on this narrative, arguing that in the cases where the use of generative AI is not clearly disclosed (how, when and where generative AI is employed), audiences are more prone to misattribute stylistic and narrative decisions to human creators, misjudge authenticity and credibility of content, and make decisions based on incomplete or misleading information. These tensions motivate the implementation of AI-disclosure labels as a way of clarifying the role of generative AI in creative outputs.

However, recent work suggests that making AI involvement visible affects how responsibility is assigned. Earp et al. (2024) argue that when AI involvement is salient, "perceived credit asymmetry" is more likely to occur, since audiences tend to attribute less moral credit to content creators, and blame for harmful content increases. This asymmetry is not discourages content creators for acknowledging the use of AI since admitting reliance on AI can reduce perceived creative achievemen (Earp et al., 2024).

These tensions form part of what van Dijck (2013) calls a reputationall economy witin the "culture of connectivity" of the social media ecosystem, where creative works are dependand of algorithms and law . The author argues that this is a matter of debate around "what counts as content, who owns it and controls it", in an environment where social dynamics are mediated by status and visibility (p.35). Therefore, creators' reputation becomes subject of popularity metrics and classifications of content and users (p.35).

Overall, ethical principles justify disclosure as a credibility-preserving intervention; governance models and regulation operationalize those principles into enforceable transparency obligations. However, implementing disclosure on platforms requires definitional decisions about what exactly counts as AI involvement and what social meaning the disclosure will produce. Legal frameworks allocate rights and accountability, but creative agency in AI-mediated production is often a dichotomical challenge. Because platforms communicate these conditions primarily through interface cues, disclosure labels function less as legal instruments and more as attribution signals that shape perceived credit and responsibility. As a result, labeling is simultaneously a transparency mechanism and a reputational intervention within platform economies, producing downstream effects on trust, credit assignment, and creator incentives.

2.5 Summary Chapter

In summary, ethical obligations (accuracy, transparency), legal mandates (EU AI Act), industry efforts (PAI), and ambiguities in authorship and copyright show why accurate labeling is a necessary mechanism within the platform ecosystem. This review on regulatory frameworks and voluntary standards, highlights the importance of labels being adaptable to context, role and creative control, and must enforce the value of transparency through accurate provenance disclosure.

Current regulation, such as the Article 50 of EU AI Act, requires disclosure solely when content is fully synthetic or manipulated, without accounting for the hybrid spectrum. As a result, labels are meant to clarify provenance and responsibility, but AI disrupts stable notions of authorship.

3 COGNITIVE FOUNDATIONS OF PERCEPTION

“The act of reading or interpretation requires an active process of constructing and negotiating meaning” (French, 2025, para. 5).

This chapter describes what happens when users encounter labeled content; this chapter explains the challenges in human perception. According to Jung et al. (2025), social media users’ perception on AI-labeled content is shaped by their mental models. For content creators, labels anticipate how audiences will perceive labels embedded in their work (Burrus et al., 2024).

3.1 Perception and its importance

The Interaction Design Foundation (2020) defines “perception as the process of interpreting information to form a mental model of the external world”.

Gibson’s Affordance Theory expands this view, describing perception as designed for action. The environment offers “perceivable possibilities for actions”, which are called affordances. Affordances are not only physical. In digital environments, affordances appear digitally.

These affordances are conveyed through the senses and are represented by cues (called signifiers). These cues can be categorized based on their cognitive demand (Ware, 2021, pp. 8-12):

Sensory Cues: Are perceived by our brain without learning and can transcend cultural boundaries. For example, the shape of a dog can be immediately linked to the concept of a dog.

Arbitrary Cues: Arbitrary cues are socially and culturally constructed representations, and therefore, must be learned. For example, the English word ‘dog’ holds meaning for those who have learned its symbolic representation.

While perception of the environment is direct, computer graphics and interfaces are indirect since they represent digital representations of objects and data. Therefore, this indirection

represents a challenge for social media platform designers since critical information must be communicated quickly and effectively through adequate cues (Ware, 2021, pp. 36-46).

3.2 Core Mechanisms of Perception

To understand how cues are perceived, Ware (2021) reinterprets Gibson's Theory from visualization's perspective, defining three sequential stages in the process of visual perception:

Stage 1:

Preattentive processing. This stage involves the immediate detection of basic visual features that stand out easily, such as color, shape and motion. Because of this, sensory symbols are easier to perceive than arbitrary symbols. This principle is relevant in interface design, where choosing the right features is key for guiding user attention effectively and immediately.

Stage 2:

Pattern organization. In this stage, the brain understands and organizes these features into meaningful and complex patterns. Here is when Gestalt principles explain how visual elements group into patterns. For instance, when applied to the context of AI disclosure labels in social media, Gestalt principles highlight the importance of placement (proximity) in relation to the post, and contrast (figure-ground) in relation to the image; a label must be differentiated from the content yet remain visually integrated in the composition.

Stage 3:

Selective Attention and Memory. Finally, only a few patterns are held in memory as the viewer holds selective queries of information, such as shortcuts, which determine what is inspected and, in turn, what interpretation can be supported by the available evidence.

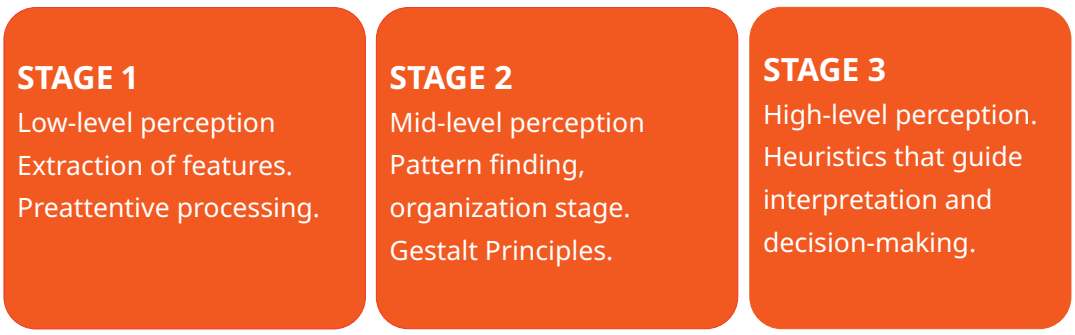


FIGURE 1.3.2. Scheme of perception as a bottom-up mechanism from Stage 1 to Stage 3 (based on Ware, 2003).

These three stages, as explained in Figure 2, describe how raw sensory information is organized and selectively stored in a bottom-up process. For an AI disclosure label to be effective, its design elements must first successfully capture attention through preattentive features (Stage 1), then adequately organised within the post layout (Stage 2), before an user can proceed to attach meaning (Stage 3).

3.3 Internal and external factors of Perception

At Stage 3, the individual uses the prior knowledge obtained from the sensory input and faces the choice of how to engage with the environment. As humans, we interpret this information through shortcuts or heuristics, and therefore prioritize information that is easy to assimilate. Perception becomes active and engagement occurs. At this point, the perceptual process pivots from bottom-top to top-down processing (Ware, 2021).

Perception at this stage is shaped by both internal and external conditions that frame and influence perception. Internal factors refer to attitudes, motives and expectations – which help guide interpretation (Robbins, 1991). External factors refer to the environment, such as the social situation and its context, which frame how interpretation is understood (Robbins, 1991).

On social media, these internal factors have direct implications for how AI labels are read. An 'AI-generated' label is not a neutral cue with a single meaning: each viewer interprets it through their own assumptions and attitudes. Jung et al. (2025) show that users hold diverse mental models towards disclosure labels depending on the type of content; from seeing them as reassurance tools for transparency to assuming that they signal synthetic content. According to Burrus et al. (2022), internal factors for content creators are the motivation to be transparent, fair, while preserving their ownership. The authors found that many creators want to use labels to demonstrate the labor behind their work and to differentiate what is 'handmade' from AI-assisted or AI-enhanced content.

External factors in social media include the content type (e.g., serious vs. humorous) and the account that posted it. The viral "Pope Drip" meme exemplifies this interplay between external context and subjectivity about what is authentic or not. This shows that the same image leads to very different perceptions of authenticity (Radivojevic et al., 2024). This perceptual instability means that users cannot reliably trust visual cues by themselves to determine authenticity, forcing them to rely on secondary learned cues such as labels, captions and the

perceived credibility of the source.

3.4 Linguistic framing as a perceptual cue

Beyond visuals, language and framing become the medium through which context is delivered and subjectivity is shaped. The framing effect demonstrates that the same factual information can seem reassuring or alarming depending on how it is presented. For instance, misleading information (such as fake news, deepfakes) is framed in a positive light, this can become more attractive. For instance, this effect can have an influence on public opinion, since our choices are influenced by the way information is framed through different phrases (Pilat & Krastev, 2025).

Figure 3.4. Framework effect in AI disclosure labels. Both posts show the same content and labels signaling AI involvement. However, the left label “Deepkafe” is designed to present AI as a warning. The right label “AI-generated content”, is neutral. These examples illustrate how different wordings of the same fact can affect perception differently (own photograph, 2025).

In the context of labels, wording matters. For instance, warning-based labels’ objective is to denote misleading content, but the framing might evoke a negative connotation. Terms like ‘AI-generated’, ‘Altered’ are often associated with misleading content (Wittenberg et al., 2023). This is a critical consideration for designing a label that is meant to signal caution or transparency regarding altered content, while simultaneously respecting the content creator’s perceived sense of authorship and informing the audience (Burrus et al., 2024).

In the case of content creators, the top-down process of perception is tied to how AI disclosure affects perceived authorship and authenticity. Labels, therefore, can also be read as statements about who deserves credit for the work. For instance, a phrase such as 'AI-generated' may be interpreted as implying that human creativity was absent or secondary.

3.5 Limitations and Biases in Perception

To understand the indirect effects of labeling, it is crucial to elaborate on what are the limitations and biases in perception. To design an effective label that successfully counters misinformation, cognitive limitations and biases must be firstly addressed:

Banner Blindness.

This phenomenon occurs when people ignore page elements that they perceive to look like advertisements. The reason is that attention is selective. Examples are banners, pop-ups and any type of ad. In essence, people direct their attention only to the stimuli related to their goals (Broadbent, 1958; Nielsen Norman Group, 2023). Over time, this becomes a learned behavior. In fact, this neutralizes the label's function, ensuring that the user's attention is directed only to the primary, often misleading, content. Research has shown that the effects of labels in the long-term can unintendedly normalize the dissemination of misleading content (Wittenberg et al., 2023).

Implied Truth Effect.

This bias, established by Pennycook et al. (2020), suggests that information that is not tagged or labeled is taken as if it was more valid or more accurate or trustworthy than labeled information. In their study, attaching warning labels to false headlines caused unlabeled headlines to appear more credible to readers.
In the context of social media, a disclosure label has the potential to increase the perceived credibility of unlabeled content, even if the content could be untrustworthy. (Altay & Gilardi, 2024).

Cognitive Load.

Ware (2004) describes the phenomenon of tunnel vision which has been linked to cognitive load, when an individual is involved in a high-load task, the detection of objects in the periphery of the visual field decreases. This limitation has implications for AI-labels. As Gamage

et al. (2025) and The Dais Report (2025) demonstrate that disclosure labels must compete with the visual dominant stimuli of a social media feed. Moreover, a static tag placed near captions can be easily perceived as background noise, and a detailed interactive label reports the same amount of cognitive effort.

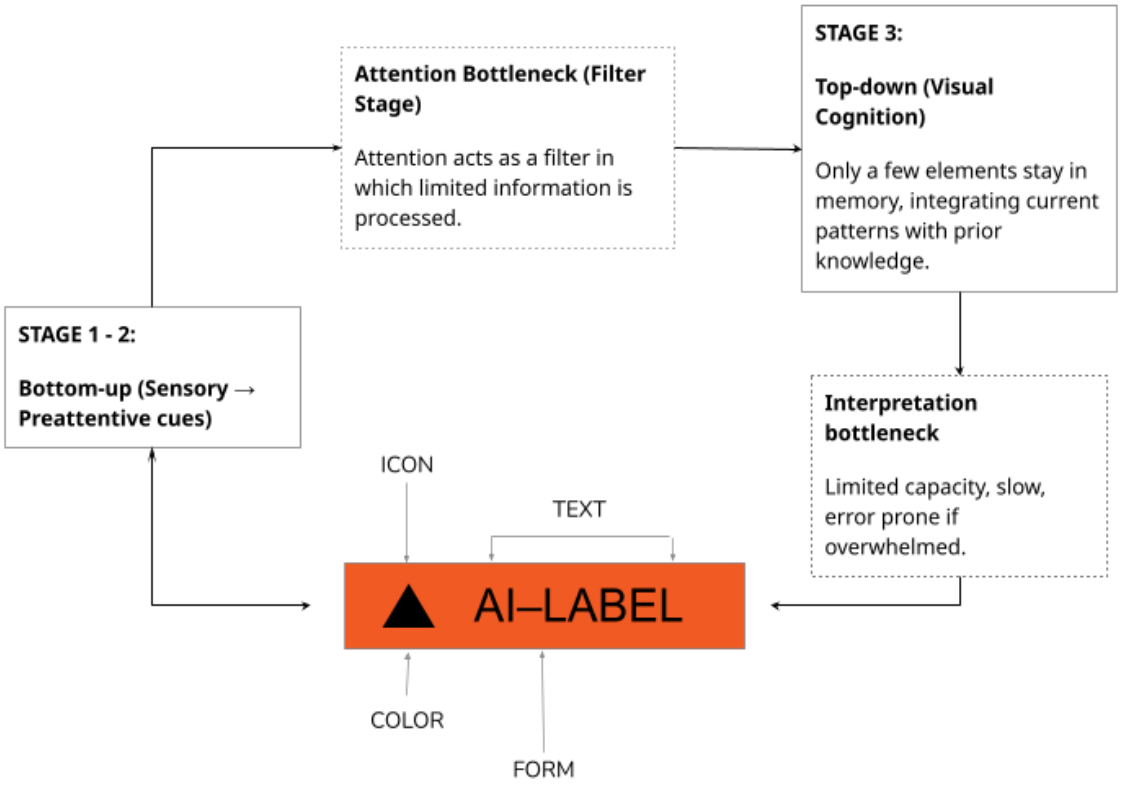


FIGURE 3.5. The Perceptual Cycle of AI Disclosure Labels. This diagram illustrates the cognitive loop based on Ware's (2004) model of visual perception. It highlights the transition from bottom-up feature extraction (Stages 1-2) to top-down cognitive interpretation (Stage 3). The markers identify critical "bottlenecks" where design-induced biases, such as banner blindness or the implied truth effect, are most likely to occur. (Re-elaboration by the author based on Ware, 2004).

3.6 Implications of AI label Design

Sundar (2008) further elaborates that the social media ecosystem represents two challenges for users: firstly, overload of information that needs constant organizing, and secondly, the lack of assurance in the legitimacy of content, which requires a continuous assessment of credibility from the user's side. Depending on how a label is designed, it can lead to positive or negative effects; this is the case with AI-labels that can have an influence in perception in persuasion contexts (Sundar, 2008). Consequently, labeling practices can have secondary effects for users' beliefs and expectations if platforms fail to implement them in a way that is

consistent with the divergent perception of users and stakeholders (Wittenberg et al., 2023; Jung et al., 2025; Gamage et al, 2025). Sundar (2008) proposes that this challenge of information overload and legitimacy (ambiguity) can be managed by effectively designing interface features that serve as informational heuristics, allowing users to make rapid automatic judgments about content trustworthiness.

Therefore, the challenge of designing and implementing labels requires a deep understanding of all actors within the ecosystem. If both content and labels are unstable in how they communicate meaning, the labeling strategy should aim to stabilize interpretation. This requires understanding not only by the social media audiences but also the experiences of content creators themselves. As van Dijck (2013) states, “actors of all kinds attribute meanings to platforms”.

3.7 Chapter summary

Perception occurs in 3 Stages from bottom-up to top-bottom. It is influenced by both external and internal factors – such as the context of the individuals – and subjective factors. For instance, social media environments interact with people through perceptual cues that act as affordances that trigger users’ mental models.
This chapter highlights different phenomena that occur when users encounter AI-generated content. Depending on how the label is presented, biases – such as the Implied Truth Effect, banner blindness and cognitive load – become side effects if the label is not implemented considering the heuristics of users’ mental models – of both audiences and creators – without causing the above mentioned side effects. It is thus relevant to consider Ware’s theory of preattentive features, Gestalt principles, and strategic linguistic framing (e.g. the weight of AI on the wording).

STATE OF THE ART

PART II

4 STATE OF THE ART

This chapter examines how AI disclosure labels are not solely policy instruments but also design features whose wording, form and placement affect how they are perceived. It reviews how prior research has examined AI disclosure labels from design, perception and usability perspectives.

Firstly, it synthesises empirical studies on label design and effectiveness, including studies on warning-label formats and engagement effects (e.g., Gamage et al., 2025) and on creators' needs, authorship concerns and perception around AI-mediated content (Burrus et al., 2024; Jung et al., 2025). Particular attention is given to how labels affect both general audiences and, especially, content creators' perceptions of authenticity and willingness to disclose AI use. Secondly, the chapter reviews provenance-based frameworks and technical standards (such as C2PA and initiatives like ailabels.org) propose concrete implementations for AI disclosure.

4.1 Conceptualising AI involvement

In 1978, Sheridan and Verplank defined the 10 levels of automation, that defined in each stage the level of AI contribution from full human contribution to full machine automation. At the lower levels (2-3), the human performs the task while the computer offers no assistance. In the 'hybrid' middle levels (4-6), the computer suggests options or executes tasks uniquely upon human approval (Sheridan & Verplank, 1978). This concept is illustrated in Figure 2.1.

Later, Parasuraman, Sheridan, and Wickens (2000) proposed a model that acknowledges human cognitive processing within the automation tasks across 4 distinctive stages :

1) information acquisition
2) information analysis
3) decision and action selection
4) action implementation.

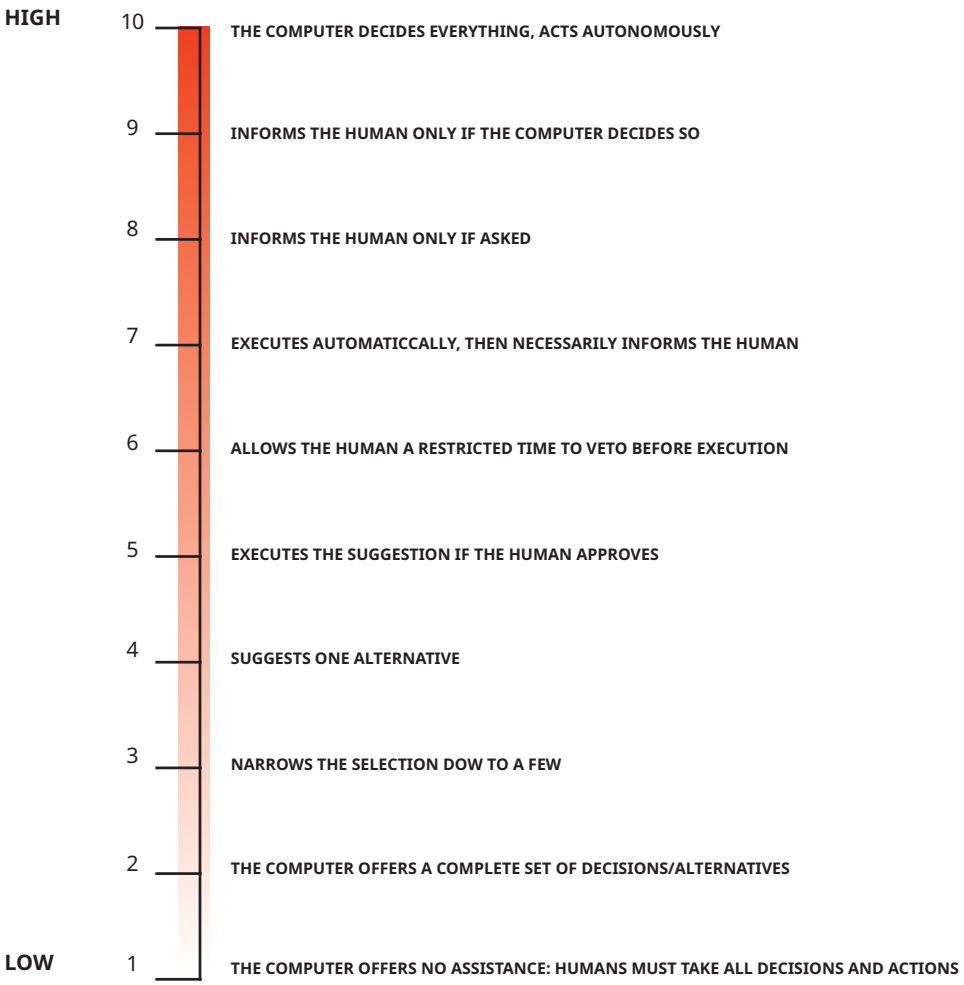


Figure 4.1. 10 levels of Automation. Adapted from Sheridan and Verplank (1978).

Similarly, each stage considers 10 stages of automation of Sheridan & Verplank (1978). Pacaux-Lemoine et al. (2011) synthesized this information into a framework that distinguishes between an agent's know-how (ability to perform a task) and their know-how-to-cooperate (ability to manage interference and share goals). Furthermore, they layered this cooperation across three levels: Strategic (mission definition), Tactical (plan carrying out), and Operational (task execution). In an assistance model, the tool merely executes a command. In a cooperation model, tasks are dynamically allocated based on criteria such as reliability and workload. Crucially, they found that while machines can effectively take over the Operational level (e.g., execution or generation), the human operator must retain control at the Strategic level to ensure the mission's intent is preserved.

As Rezwana & Maher (2022) present is a framework for modeling interaction between humans and AI in co-creative systems (COFI) that highlights interaction components such as participation style, contribution type, and communication. From Rezwana & Maher (2022), one can observe a recurring pattern of human-AI involvement in co-creative systems:

1) Definition: Setting the intent, constraints, defining character attributes and style rules before or alongside making outputs.
2) Generation: The AI generates outputs within the conceptual description of the first stage.
3) Selection: The user "selects from AI-provided options" (curation), which is how direction gets enforced after generation.
4) Evaluation: Evaluation can be done by the AI (as evaluator) and it commonly leads into refinement (polishing).

4.2 Disclosure design patterns

Gamage et al. (2025) state that current labelling practices lack grounding investigation about the nuance of synthetic content that must be considered when designing for AI content. Therefore, they propose a design space for labels that are applied to AI-generated content on social media, and offer an evaluation on their efficacy and impact on user perception. To clarify, the selected study solely evaluates labels as design interventions, but does not dive deeper into content creator's perception or how these labels represent the true nature of workflows or the spectrum of AI involvement. . However, Gamage et al. (2025)'s study offers a foundation for label design that could further support further studies on content creators.

In their experiment, participants saw 10 social media posts with different label designs on a social media simulated feed, applied to political and entertainment images (of AI generated and real provenance). The design prototypes were studied across 4 design dimensions within the controlled experiment: (a) Wording, (b) color and iconography, (c) position, (d) level of detail. Each element was found to have an impact on trust, comprehension and engagement.

(a) Label wording. Gamage et al. (2025) define it as a critical dimension of AI–disclosure labels design. It is about how the label is framed, since a short phrase does not only state AI involvement, but how much AI could have intervened in the creation of the content. The more accurate a label seems, the higher the level of comprehension about the legitimacy of the content. The authors found how the framing of the label indirectly affects the perceived trust of the content and the level of engagement (commenting/sharing/liking).

From a design perspective, Gamage et al.'s study suggests that wording holds most of the interpretive weight of AI labels. Detailed effects of labels on perception are considered in Section 4.6. In this section, the point is to explain why linguistic differences in labels are important as part of the design components of a label, as they aid users in understanding the provenance of the content generation.

(b) Color and icons influence perception. This dimension mainly modulates the tone and salience of the wording. Gamage et al. (2025) tested two types of icons: warning icon (<!>) and neutral icon. The results show that neutral icons and colors were preferred and perceived as more trustworthy. Conversely, red icons with a hazardous tone consistently decreased trust in the content. Interestingly, the lack or presence of an icon in terms of preference was not relevant. The presence or absence of the icon had neutral effects, in the sense that it mattered less.

(c) Position. This dimension focuses on the placement of the label on a post. The authors tested how intrusive or acceptable labels felt to participants. In terms of position, Gamage et al. (2025) introduce three types of placements:

(1) Label on content (obscuring): Covering part of the content.
(2) Proximity placement: Above or near the content without overlap.
(3) Integrated on content (non-oscuring): Embedded on the image.

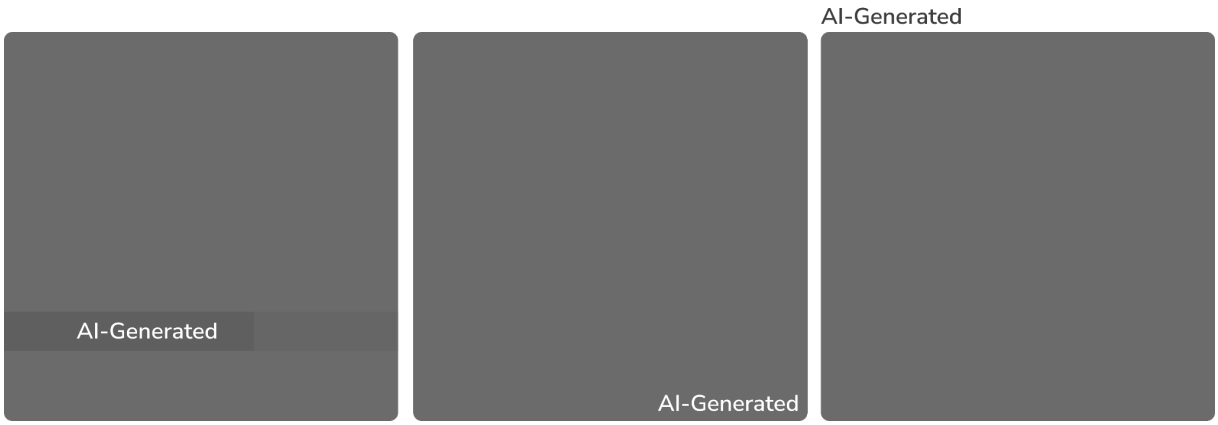


Figure 4.2a. The first image illustrates placement type (1) label on content (obscuring), the second image illustrates placement type (2) proximity placement, and the third image illustrates placement type (3) integrated on content (non-obscuring). Adapted from Gamage et al. (2025).

(d) Level of detail. Gamage et al. (2025) describe this dimension as the amount of information a label reveals in order to provide context to educate the user and promote transparency (in order to consider regulatory measures). The level of detail can range from a simple label with a single text, to detailed information. On a simple level, it consists of showing concise information that is fast and easy to read in a social media feed. And on a nuanced level, it refers to labels that offer more granular explanations about types of AI involvement. Systems such as Content Credentials can enable granular disclosure, but that depends on how the platform decides to display the information.

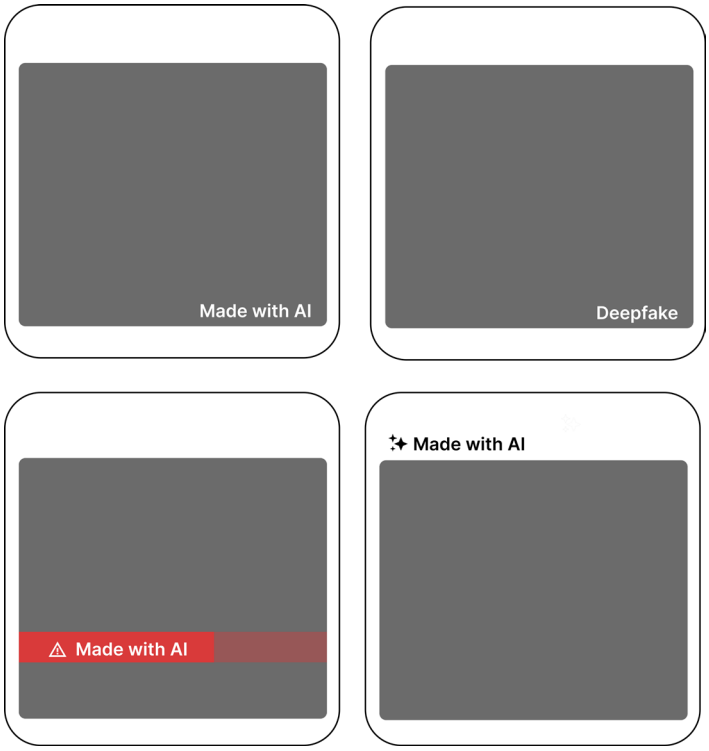


Figure 4.2b. Recreated illustration of a sample of 10 different prototypes across four dimensions (wording, color and iconography, position, level of detail (Content Credentials)) based on Gamage et al. (2025).

Overall, Gamage et al. (2025) 's results emphasize that wording and tone shape are essential in shaping trust and comprehension, while visual style and placement reinforces their impact.

4.3 Progressive Disclosure & Provenance

Gamage et al. (2025) mentions progressive disclosure and provenance as the fourth dimension a label can reach. This section explores the technique of provenance, employed by C2PA.

The Interaction Design Foundation defines progressive disclosure as "a user experience (UX) technique that defers advanced features and information to secondary user interface (UI) components". The idea is to maintain relevant content in the main user interface and make advanced content accessible underneath the layers upon request. It seeks to provide users with what they require at the appropriate time (IxDF – Interaction Design Foundation, 2016).

The C2PA (Coalition for Content Provenance and Authenticity, 2025) is a non-profit project from the U.S.A that focuses on developing technical global standards for content disclosure. Their aim is to establish an ecosystem of digital provenance in industry platforms (such as

editing tools (Adobe Creative Cloud) and social media) by revealing provenance data of the content in terms of depth and breadth.

C2PA's approach uses a multi-layered structure:

(1) Level 1: Title

(2) Level 2: Signer and date

(3 Level 3: Content summary

(4) Level 4: Creation info

(5) Level 5: Link to full view

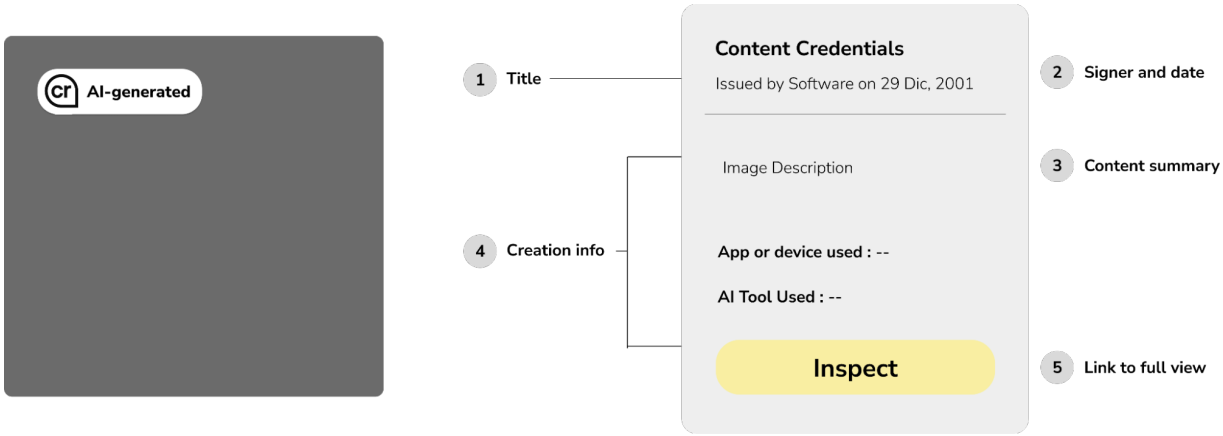


Figure 4.3. Representation of how C2PA (Content Credentials) could appear through progressive disclosure in a social media post. Adapted from C2PA documentation Burrus et al. (2025), p.41.

Recent experimental work has found that C2PA can, in fact, increase perceived credibility by prioritizing provenance transparency of the content (C. Trattner et al., 2025). Similarly, He et al. (2025) introduced the concept of granularity for AI-disclosure practices and how this approach should involve how AI was employed in detail. However, both Burrus et al. (2024) and Gamage et al. (2025) found that a more nuanced technique could cause high cognitive load on users, considering that social media interfaces are made for rapid scrolling.

4.4 Typologies of AI labels

In addition, Wittenberg et al. (2023) distinguishes labels in two main categories:

1. **Process-based labels.** This category refers to labels that signal how the content was made. It focuses on providing transparency about content authorship. Examples are la-

bels phrased as 'AI-generated', 'AI-assisted', 'AI-enhanced', or, provenance-based badges (e.g., Content Credentials) that centers on progressive disclosure to show more information.

2. **Warning-based labels.** This category refers to labels that are concerned with the potential consequence of viewing the content. Their focus is on the potential harm or misleading intent associated with the content. Examples are labels phrased as 'Deepfake', 'Fake', 'Synthetic' or 'Misleading'.

Gamage et al. (2025) note that process-based labels often carry a neutral connotation: a label phrased 'AI-generated' can appear on an abstract art or to a critical news post. By contrast, impact-based labels are more concerned about warning the user, by immediately flagging the content as potentially deceptive or harmful.

Beyond these categories, some emerging frameworks have started to focus on more granular and creator-friendly approaches to labeling practices. Ailabels.org, an open source initiative by Zach Rattner (n.d.), proposes a voluntary framework for content creators to disclose AI. Its goals are to aid content creators communicate how AI contributed to their work and to promote ethical transparency through self-disclosure.

Figure 4.4a. AI Labels categorization as presented in the official specification.
From AI Labels: A New Voluntary Content Framework, by Z. Rattner, n.d., AIlabels.org (https://ailabels.org). CC BY 4.0.

Although not adopted on social media, the framework serves as an insightful approach towards creator-centered label design.

As figure 4.4a shows, "Made by Humans with AI", the human is framed as the main contributor and AI as an extension. On the left, there is a fingerprint icon sitting above a small processor icon containing the letters "AI". On the right, the phrase "Made by Humans" and then "With AI" structures a hierarchy: human first,then AI as an addition. Underneath, the description "Human Animation, AI Storyboard", frames AI as responsible for a specific sub-task. In the second label, "Made Primarily by AI", the distribution states AI as the main producer. Beneath, the description adds "AI Synthesis, Human Retouching", human work is framed as corrective.

Human attribution and co-partnership with AI

Human-AI co-creation is an emerging field, while AI-assistance can help support or enhance skills, co-creation or collaborative contribution blur the boundaries of who is the author and who deserves attribution of the work, He et al. (2025) identified three dimensions of creative attribution: contribution type, amount of contribution and initiative.

Accoridng to the findings of He et al. (2025), audiences perceive authorshop credit depending on the nature of AI-contribution. Therefore, contribution is measured within a spectrum: “contributions of different types warranted different levels of authorship credit”. Additionally, He et al (2025) crafted an AI attribution statement based on the research and translated it into UI parameters:

Figure 4.41b. Anatomy of a granular attribution statement showing attribution components (amount, contribution type, initiative, review status, and tool/version).Reprinted from IBM and He et al., (2025) research https://aiattribution.github.io/

However, He et al. (2025)’s participants within the study were employees for an international technology company, and different viewpoints are needed. The authors emphasize on examining more complex co-creative workflows as a future work recommendation. For AI disclosure, the authors suggest a granular approach; this can complement C2PA, considering the principle of progressive disclosure to show the extent of AI contribution.

4.5 Current Industry Labeling Practices

This section reviews current platform labels in terms of how they practice AI disclosure, and compares the approaches of four key platforms – Meta, TikTok, Youtube and ArtStation – across three dimensions: scope, mandate, and label framing.

Meta (Facebook,Instagram and Threads) currently approaches self disclosure and platform automated detection. It primarily targets synthetic and heavily edited video content and is

gradually beginning to implement C2PA industry standards. Within Meta's governance structure, the Oversight Board functions as an independent review body that evaluates controversial platform decisions and issues that affect the public.

Meta began in 2024 to apply labels to videos, audio and image content. The labels are displayed as an 'AI info' tag on top of content posts on Facebook, Instagram and Threads. Future plans include utilizing more nuanced terms – such as 'Made with AI' or 'Imagined with AI' for photorealistic images (Bickert, 2024).

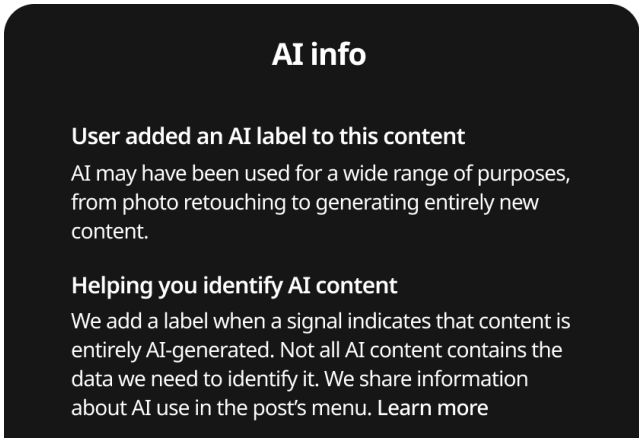


Figure 4.5a. Example of Meta's AI-generated content label (adapted from Instagram, 2025).

TikTok applies AI-labels on fully synthetic content. Simoultaneously, the platform encourages content creators, to disclose AI content manually (TikTok, 2023). Their policies are based on the study of Wittenberg et al. (2023) who researches viewers' perspectives regarding different label styles. For instance, two types of levels are currently utilized:

The first label, 'Creator labeled as AI-generated', indicates whether the content is fully synthetic or significantly edited with AI. The second label, 'AI-generated', is automatically applied if the platform identifies that the content as completely generated with AI; this occurs when a creator uses Tiktok AI effects or external tools. However, if the platform detects synthetic content that is not self-disclosured, removal of it and restriction to the account might be applied.

TikTok (2025) currently differentiates synthetic content in two categories:

- Creator labeled as AI-generated. This type of label is used by the content creator.
- Significantly Edited Content. Platform automated label.

In addition, TikTok is one of the first platforms to implement C2PA-based provenance on content. Firstly, Content Credentials are attached to images and videos in order to autolabel

them as ‘AI-generated’ when such metadata is present in the content (TikTok, 2024).

Figure 4.5b. Example of TikTok’s AI-generated content label. (Adapted from TikTok, 2025).

Google (YouTube). Labelling is presented as a way of adding context so that audiences understand the intent and background of videos. For instance, one type of label is used: ‘altered or synthetic content’ for both videos and reels, if the content is meaningfully AI generated. In contrast, unrealistic or minorly edited content such as beauty filters or enhancing effects won’t be required to be disclosed (TeamYouTube, 2024).

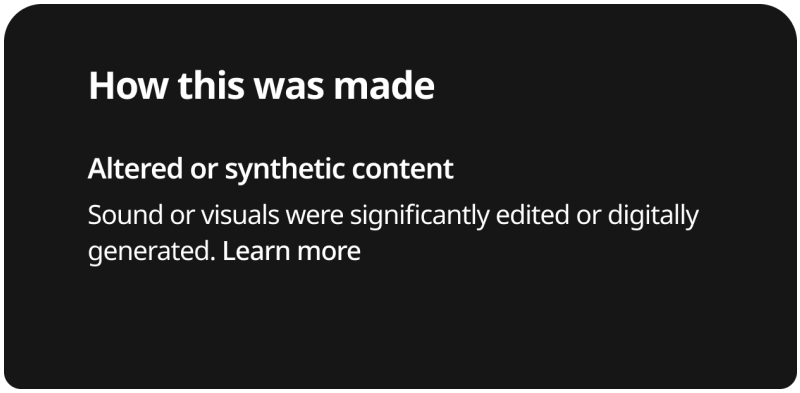


Figure 4.5c. Example of a “How this was made” panel indicating altered or synthetic content (author’s illustration based on YouTube’s interface, 2024).

A**rtStation** has adopted a different creator-centered approach. Instead of requiring creators to label AI-generated content in the feed, it offers an optional “#NoAI” tag that adds an HTML meta tag to forbid the use of tagged work for training AI models. This approach aims to give artists agency. AI-assisted artwork is allowed on the platform, and transparency about tools and process is encouraged, yet they do not enforce explicit penalties.

Figure 4.5.d. Example of ArtStation disclosure

binary hashtags.

Overall, Meta, TikTok, Youtube and ArtStation do not share standard visual languages on labels, and their policies are implemented differently. Lindsey, D. (2025) argues that platforms implement rules differently: Meta, Youtube, TikTok, ArtStation, combine creator-side disclosure (using hashtags or label toggles) with automatic detection tool, and differ in how they apply penalties to users who do not follow these policies.

Platform	Label used	Mandatory disclosure	Penalties	Who Applies the Label?
Meta	AI Info	yes	yes	Creator + Platform
YouTube	Altered or synthetic content	yes	yes	Creator + Platform
TikTok	AI-generated	yes	yes	Creator + Platform
ArtStation	##noAI, ##AI (or related)	–	No explicit penalties	Creator or Platform

Figure 4.5.e. Summary table of AI-related labeling across platforms describing labels used, and whether they impose mandatory enforcements.

4.6 Audience Perception Studies

This section discusses how different labels are perceived in contrast to the context of the social media content (politics, news, art, or marketing). As Burrus et al.(2024) found, labels influence how the content and the creator are perceived depending on the content type.

News, politics contexts

Burrus et al. (2024) found that when audiences encounter AI labels on news content, they tend to interpret the information as less trustworthy. In particular, the 'AI-generated' immediately triggers skepticism towards the legitimacy of the content. Atlay and Gilardi (2024) similarly show that the label reduces belief in the content by approximately 2.7 points, reflecting audiences assumption of full AI contribution.

A label that provoked even more skepticism on the content – but for a different reason –was 'Deepfake'. The underlying reason is because it holds alarmist connotations and felt ambigu-

ous to participants (Gamage et al., 2025).

Fashiont, entertainment, marketing, comedy contexts

In comedy/humorous contexts, audiences find AI use as more acceptable and labels "tend to have less impact on perceptions of authenticity", since users' judgment of credibility is dependent on diversion or joy (Burrus et al., 2024).

In contrast, within lifestyle influencers or photography domains, fully AI-generated content is perceived as inauthentic because users' judgement on credibility is dependent on human effort. This means that when creative content is part of self expression or leads to monetary or social capital, it is accepted if labels imply minimal AI involvement (Jung et al., 2025; Rae 2024).

In conclusion, in low stakes, minimal AI contribution is often accepted, however fully AI-contribution raises authenticity concerns and users doubt the effort (Jung et al., 2025).

In conclusion, audience perceptions centre on legitimacy/trustworthiness, satisfaction, authenticity and perceived effort. And aside from the context of the content, and in broad terms – Jung et al. (2025) and Rae (2024) found that audiences prefer terms that are both accurate and clear about AI involvement:

'AI-modified': Implied that there was human input, but AI modified the content.
'AI-generated' or 'AI-created': Implied clarity and neutrality regarding full AI involvement.

Least preferred labels:

'AI Info' and 'Made with AI': Perceived as ambiguous and did not provide clear information about the extent of human contribution.
Labels implying AI assistance: Were perceived as less accurate.

Overall, users prefer clear and accurate labels. In high-stakes contexts and in low-stake contexts that involve creative expression or monetary credit, values such as transparency are appreciated, and for that reason accurate and neutral labels are more valued. However, current label strategies are binary: AI or no AI. This can have an indirect impact on content creators. According to audiences, the degree of AI involvement is equivalent to the lack/amount of human effort from the creator's side.

Rae (2024) highlights that the relationship between audiences and creators is critical for a

creator's reputation. In these cases, voluntary self-disclosure can represent a risk. Burrus et al. (2023) and Jung et al. (2025) propose that, if viewer's perceptions change depending on the content type and the account that posts it, the information the label represents must adapt accordingly.

4.7 Creator Perceptions of AI, disclosure and platforms

What research says about creators' challenges and motivations

Content creators experience labels as signals that can affect their perceived authorship, originality, reputation and visibility within the platforms. Existing research shows that creators' perceptions towards AI labeling practices can be helpful when the intent is transparency; however, they may feel unfair or stigmatized if they force hybrid practices into a single binary category (Burrus et al., 2024; Wittenberg et al., 2025; He et al., 2025).

On one hand, labels can carry the risk of stigma, since AI-generated work is associated with controversies concerning whether content is real or trustworthy. In fact, mandatory labeling has the potential to erode the sense of creative ownership by requiring creators to reveal the use of a tool that is currently central to public debates around copyright and originality (Burrus et al., 2024).

In fact, labeling practices on social media can influence a creator's public image and reputation since they have an impact on perceived authenticity and creative ownership. In fact, the choice to disclose AI use in social media forces content creators to negotiate between maintaining authenticity and optimizing for performance (Burrus et al., 2024).

On the other hand, disclosure can be an opportunity for creators who use AI. Burrus et al. (2024) show that creators who work with AI tools prefer labels that explain how AI contributed - whether it generated ideas, edited existing material or assisted in production; while creators who avoid AI are interested in "hand-made" or "no-AI" labels that distinguish purely human work.

What creators say in online discussions

To complement academic studies, I conducted a small-scale digital ethnography on Reddit, examining discussions in three threads where content creators debated about AI and labels in two subreddits r/aiwars and r/NewTubers. This intends to be an exploratory reading of threads and comments to identify recurring concerns of creators.

Frustrations with inconsistent labeling practices. Creators expressed that only a minority of AI use is visibly labelled, while AI-assisted content is not disclosed. One content creator complained that “only a minority of AI-use gets labelled and singled out”, noting that “there should be way more AI labels” (Reddit user, 2025).

Uncertainty and frustration about when AI should be labelled. Here creators comment about YouTube and other video and music platforms; they questioned whether minor uses of AI tools, such as automated mastering plugins, color correction or resizing in video editors, should show the same label mentioning ‘AI’ in the content: “if you use AI video tools such as color correction or image resizing, then you must flag your video as using AI, or be in violation of Google’s rules “ (Reddit user, 2025). In fact, hybrid workflows were perceived as difficult to label or classify. Many worried that broad or vague rules would force them to label almost everything as AI-related or to hide AI use entirely: “Overly vague laws coupled with fanatical AI hatred makes the most likely outcome a complete lack of disclosure. This would lead to a bunch of tools created to “prove” AI use” (Reddit user, 2025).

Concerns about stigma and harassment associated with AI labels. Creators on reddit anticipated feeling attacked for disclosing. They expressed how AI labels made the audience assume that AI work involves no effort and offered no legal protection. One creator noted that people “get harassed for no other reason than utilising AI, or even being suspected of using it” (Reddit user, 2024).

Debates about the legal scope of AI-labels. On the NewTubers community, creators asked how far mandatory AI labels should extend. A participant asked if AI labels would apply to all content using AI or including AI-assisted art or only to media that could be used for misinformation. Some participants expressed that platforms would “label everything as AI” to avoid legal risk.

What creators do on social media platforms

This section focuses on observing how content creators engage with labels or disclose their use of AI on YouTube and Instagram.

On YouTube, content creators add disclaimers on descriptions under the video. In the case of fully synthetic content, the creator will add it as well. However, when the content is AI-Assisted, the creator will mention the tools used on the description or on the title by default (or through hashtags). The platform by itself enforces strict disclosure, and the most com-

mon method of self disclosure is by including the word 'AI' next to the title description. Most content creators that use fully synthetic AI for their content will mention it in their biography as a description and most of the time, their content is for experimental purposes. Often, the content creators that are open about their use of AI are mostly generative artists. Other creators that are open about their AI use focus on creating fake personas or showing the audience how to create them.

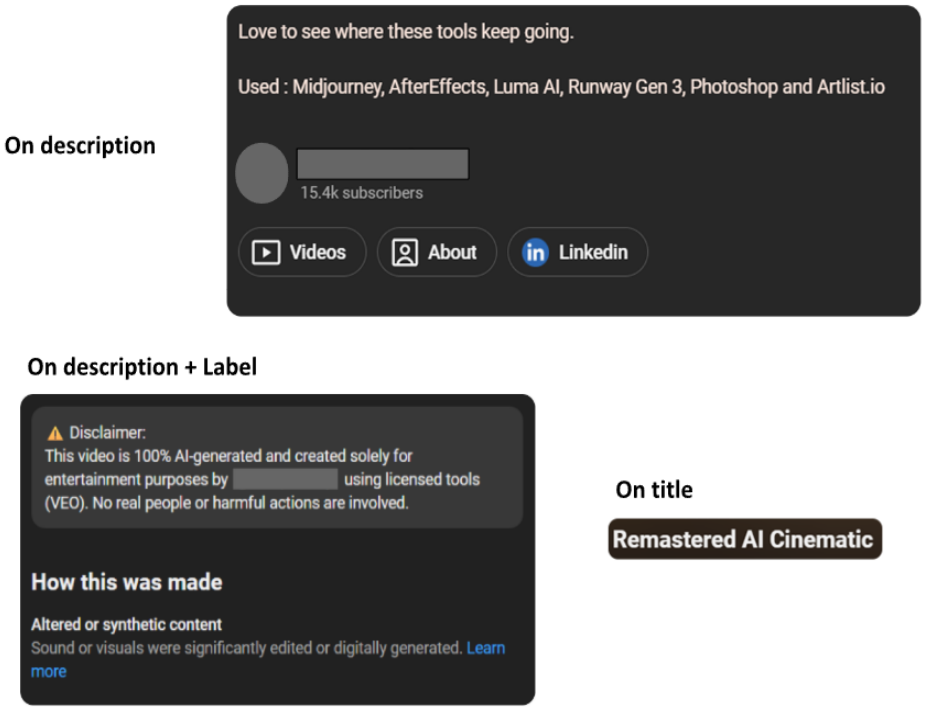


Figure 4.7a. Examples of how AI use can be disclosed on YouTube: in the channel description, via the platform's "AI-generated" label and "How this was made" panel, and in the video title (author's illustration based on YouTube interface, 2025).

On Instagram, content creators, such as comedians do not disclose their use of AI when they employ filters. In contrast, profiles of photographers who use AI for edition update it as a description under the content and add hashtags. Other artists who claim full use of AI, are open to disclose it in their biographies. Another method creators use is to add the word 'AI' next to their username. Mostly when humorous fully synthetic content could be misinterpreted, the platform or account will add the 'AI Info' label.

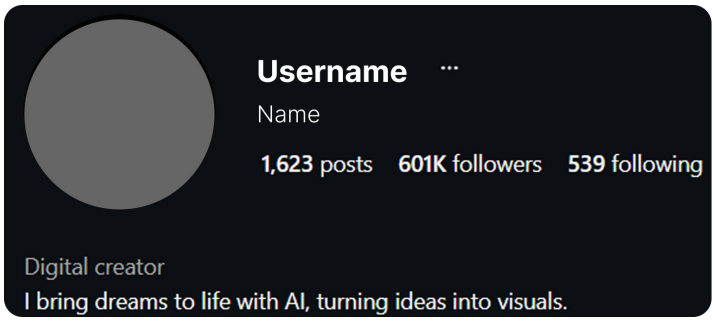


Figure 4.7b. Example of Instagram profile where AI use is signalled in the bio ("Digital creator… I bring dreams to life with AI, turning ideas into visuals") (author's illustration based on Instagram interface, anonymised).

On TikTok, some content creators only focus on uploading fully synthetic content for entertainment purposes and actively label their content under the label 'creator labeled as AI-generated'. For example, when they upload a video of a political figure on a dancer's body, they assign the label and write a disclaimer on the captions.

However, this is not always the case; when the content does not involve a public figure or contains sensitive elements, fully synthetic-humorous media is often not disclosed. Another fact is that the majority of users within this platform are influencers who engage in passive disclosure – especially when they employ filters– that are automatically disclosed by the platform under the label "filter effect", which appears to be a new, recent form of specified tag (view Figure 4.7c).

Figure 4.7c. Example of an AI-disclosure filter label prototype in a TikTok-inspired interface (author's simulation, not an actual platform UI).

Overall, across the three platforms, content creators use titles, descriptions, and even in their own usernames to inform the audiences about their use of AI; some disclose it in their profile/biographies as part of their identities. For example, the profile of the "experimentalist" –who uses AI tools to create innovative content– ,or the "generative artist" who openly creates AI Art. Other accounts that use humorous content that is very obvious to be AI –or does not contain elements that could be misinterpreted, are usually not labeled regardless of their AI provenance. However, among the stricter platforms, YouTube is the one that avoids unlabelled synthetic content the most, and even then, creators will either employ the word AI (in the title of the video or reel) to attract audiences interested in the technology, or to comply with platform disclosure rules.

Another general observation is that disclosure is contextual. On one hand, disclosure seems

as an unnecessary activity to comedians or humorous, specifically, when the syntheticity of a filter or generated content is irrelevant to the goal (making the audience laugh). On the other hand, photographers and artists inclined to hybrid or partially edited content, experience the need to distinguish their work by detailing the reason and objective AI is used, either in their biography, through hashtags or captions.

4.8 Synthesis of Gaps

In low stakes contexts where the value is not objective truth, and where the human intent and expression are more valued (e.g., artistic contexts), current disclosure practices often carry a social and reputational risk: the creator's perceived effort and authenticity are compromised, hence their needs need to be evaluated in terms of their role and effort (Burrus et al, 2024; Rae, 2024; He et al., 2025).

Labels are categorized as process-based or warning-based (Wittenberg et al., 2023. Although these categories differ in intent, platforms often implement both as binary tags in the initial interface layer (He et al., 2025). Therefore, creators desire more descriptive process-based labels that explain the AI's role (ideas, assistance, editing) (Burrus et al., 2024). In practice, creators manage to self-correct the binary nature of labels: by using titles, descriptions and hashtags to create their own narratives surrounding the use of AI. This implies that creators use other solutions to provide the detail and explanation their audience demands that the platforms cannot provide, regardless of the various strategies provided in an attempt to balance regulatory mandates and creator's needs through creator-self disclosure labels and platform automated labels.

Platforms such as Meta, TikTok, and YouTube currently lack a shared visual language and their reliance on binary labels ignores the nuances of co-creation (He et al., 2025). In creative domains, audiences often value effort and authenticity (Jung et al., 2025). Therefore, creators fear that binary labels erode their sense of ownership and invite social stigma (Burrus et al., 2024). This friction leads many creators to use titles and hashtags to provide the detail that current labels do not provide. The chapter identifies a gap in current practices: a need for a creator-centered approach that accounts for hybrid workflows. He et al., (2025) proposes to consider workflow disclosure and in which stage the human and the AI contributed.

A creator-centered approach that accounts for hybrid workflows and human authorship (similar to the voluntary AIlabels.org framework or He et al. (2025)'s approach of granularity) is an ongoing research aim.

4.9 Summary Chapter

AI disclosure labels function as design interventions that shape creator reputation. These labels comprise four distinct dimensions: wording, color, placement, and level of detail (Gamage et al., 2025). Meta and YouTube enforce mandatory disclosure, however these platforms rely on binary tags in the sense that they compress the complexity of co-creation (He et al., 2025). Consequently, content creators adopt manual workarounds, such as hashtags and biography descriptions, to defend their creative effort. Consequently, a gap remains in current design: the absence of process labels that are context sensitive (news, art) and consider attribution frameworks that account for the complexity of human and human-AI workflows.

THESIS IDEA AND PROPOSITION

PART III

5. THESIS IDEA AND PROPOSITION

This thesis investigates AI disclosure labels on social media, focusing on the authorship-related message they convey and its effects on creators. To do so, I conducted a qualitative study with content creators who evaluated different label types across content contexts in a simulated social media feed.

Main RQ:
How do AI disclosure labels shape content creators' (a) sense of authorship and (b) willingness to disclose AI use on social media?

RQ1:
How do creators interpret what AI disclosure labels imply about who made the image (authorship and credit)?

RQ2:
Which factors influence a creator's willingness to disclose AI involvement on social media?

6 STUDY DESIGN

6.1 Overview

The study aims for a qualitative approach that combines a think-aloud method with follow-up questions. Six participants, anonymized for confidentiality, were selected based on their common denominator of being content creators that work across fields such as photography, illustration, music and 3D visual work.

During the sessions, participants explored a set of labeled images while verbalizing their interpretations in real time. The follow-up questions focused on expanding on how labels and AI shape their sense of authorship, trust and personal creative identity.
All sessions were audio-recorded, transcribed and analysed using reflexive thematic analysis

based on Braun and Clarke's framework, in order to identify patterns in participants' views on authorship, identity, and trust toward disclosure labels and AI tools.

6.2 Study Components

The empirical study addressed the following objectives:

- To investigate how AI-disclosure labels applied to AI-generated images in different contexts (news, art, interior design) shape content creators' self-perception of authorship, their perception of image realism, and their willingness to disclose their use of generative AI on social media platforms.
- To analyze how content creators interpret different AI disclosure wordings ('AI-generated', 'AI-assisted', 'Made with AI', 'Co-created with AI') depending on different images.
- To examine how label framing (wording) influences content creators' sense of authorship and their self-perception (identity as content creators).
- To identify the factors that motivate or discourage content creators from applying disclosure labels to their AI-generated or AI-edited images.

6.2.2 Participants

The study included six content producers. They were all from Latin America (Chile, Peru, and Mexico) and ranged in age from 28 to 43. Every participant shared their work on digital platforms on a regular basis and worked in creative fields either professionally or semi-professionally in creative fields and regularly shared their work on digital platforms.

Participants	Age	Role	Country	Social Media
Participant #1	42	Illustrator (Digital and traditional)	Mexico	Instagram, Facebook
Participant #2	28	llustrator (Digital)	Mexico	Instagram
Participant #3	29	3D Artist (VFX)	Mexico	Instagram, ArtStation
Participant #4	29	Creator, Influencer (Film and Writing)	Peru	Instagram, TikTok

Participant #5	33	Performer (Music)	Chile	Instagram, Youtube
Participant #6	29	Photographer (Digital)	Mexico	Instagram

Participants represent a small but diverse sample of visual and narrative creators who work with images, stories and performances in different media, and who have varying levels of familiarity with generative AI tools.

6.2.3 Materials and stimuli:

The main materials consisted of four mock social-media posts (image + caption + label) and an interview guide. Each post combined a specific AI-disclosure label with a particular visual and contextual setting. The participant interviews are in Appendix A.

6.2.4 Procedure

At the beginning of each session, participants received an information sheet describing the study's aims, procedures and data-handling practices. Written informed consent was obtained before data collection began.

Preliminary questions:

- Which social media platforms do you use the most?
- What type of content do you typically create or post?
- Have you ever created content using generative AI tools (like Midjourney, DALL-E, etc.)?
- Have you seen AI-labels on social media before?
- If yes, have you ever used them?

Think-Aloud Session:

Participants were then introduced to the think-aloud task with the following explanation:

"I'll show you a few examples of social media posts that include different labels, and I'd like you to say out loud what comes to mind as you see them, what you notice, what you think the label means, and how it makes you feel about the post. There's no right or wrong answer; I just want to understand how it feels from your perspective."

Consequently, they were presented with the four labeled images representing different contexts (news, interior design, art, speculative art) and asked to verbalise their thoughts and feelings aloud while viewing each image. The researcher only intervened with neutral prompts to maintain the flow of speech.

Follow-up interview:

- How do you feel about the existence of these AI-labels in social media?
- In what ways do you think labeling might influence your audience's trust - positively or negatively?
- What would make you more willing to disclose AI use in your content?
- What comes to your mind when you imagine seeing or using a label that says 'Human-made'?
- And lastly, has the existence of AI influenced your creative work?

This analysis followed an inductive and qualitative approach inspired by Braun and Clarke's six-step framework for reflexive thematic analysis. Following the completion of the think-aloud sessions and follow-up interviews, all audio recordings were transcribed and imported into Taguette, an open-source software for qualitative data coding of interviews. The purpose of the analysis was to understand how content creators made sense of AI disclosure labels in relation to authorship, creative identity, and credibility of the content (images). During this stage, preliminary observations were annotated. Subsequently, coding was extracted from inductive values, or main beliefs from participants' quotes.

Once initial coding was considered finalized, the codebook was exported and organised with the support of a customized digital visualization tool that I developed for this study. This tool facilitated the grouping of codes across interviews. All decisions were made manually, considering the research objectives. Codes expressing similar meanings were extracted from the Taguette exported pdf, grouped into categories and reviewed alongside the original transcripts.

The themes emerged through Braun & Clarke's reflexive thematic analysis approach, as patterns that described the shared meanings and values across participants' opinions on images with AI disclosure labels. I extracted themes from the most repeated codes. Firstly, I categorized them in identity, and all the codes that mentioned how the creators saw themselves. Secondly, I categorized by labels. Thirdly, I categorized by interpretations on platforms and social dynamics. The fourth label category was based on attitudes towards AI tools. The Thematic Analysis can be found in the link on Appendix B.

6.2.5 Probes (view appendix D)

Label: 'AI-Generated' (Context: News Post)

FIGURE 6.2.5a. A stock photo from Pexels (public domain image, cited in the appendix) was intentionally used instead of an AI-generated image. This choice is aligned with ethical practices and legal mandates in high stakes contexts. Specifically, the EU AI Act requires clear and visible labeling of synthetic media used for public interest (European Commission, 2025), but news organizations often avoid synthetic images to preserve public trust. Using an authentic stock image with 'AI-Generated' label has the objective to test the label's impact of perceived authenticity when applied incorrectly or as a misattribution, providing critical data points on public skepticisms (Gamage et al., 2025). The label is used by TikTok.

Note: The stock image was intentionally used rather than an synthetic one to reflect real world practices where authentic images are confused as fake.

Label: 'AI-Assisted' (Context: Digital Artwork)

FIGURE 6.2.5b. A AI-enhanced stock image was used to test the 'AI-assisted' label. This simulation reflects on the industrial understanding of the "assisted paradigm," where AI acts as a utility for refinement, editing for human initiated work (Rae, 2024).

Note: The 'AI-assisted' label validates the wording where AI acts as a utility for refinement or edition. Based on Night with her Train of Stars (1912) by Edward Robert Hughes. This public domain image was modified by the author to simulate AI-enhancement for the purpose of the study.

Label: ‘Made with AI’ (Context: Interior Design Render)

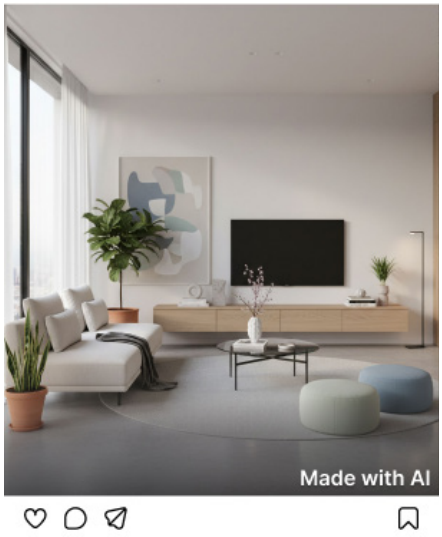


FIGURE 6.2.5c. A A fully AI-generated image was employed to represent common commercial use of generative tools, such as creating design concepts or lifestyle renders for marketing. The image’s function as a conceptual and illustrative probe to assess user interpretation of the attributional label aligns with strict academic publishing guidelines (Elsevier, n.d.; SciPub+, 2025).

Label: ‘Co-created with AI’ (Context: Speculative Art)

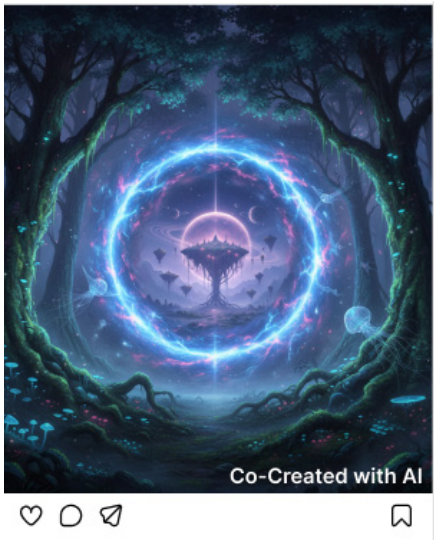


FIGURE 6.2.5d. A synthecic art fiction image functions as a design probe to test reactions to collaborative framing.

This image and label are inspired on Rezwana & Maher, (2022) studies about collaboration and partnership with AI.

7. FINDINGS

This chapter reports patterns from the content creator interviews and think-aloud sessions, using reflexive thematic analysis. The findings answer the research questions in two steps:

- How do creators interpret what AI disclosure labels imply about who made the image (authorship and credit)?
- Which factors influence a creator's willingness to disclose AI involvement on social media?

7.1 Overview of themes

The analysis answers the research questions with four themes that describe the patterns observed from the interviews Theme 1 corresponds to how participants view the concept of authorship and they defend Ideation as a primary core of their identities as content creators. Participants are visual artists, musicians, writers and photographers who create content for social media.

The following themes, Theme 2 and Theme 3, explain the patterns from the think-aloud sessions where participants interacted with each of the four images with their respective labels. Two patterns emerged among labels that were precise in what they wanted to communicate and labels that were ambiguous in what they intended about the extent of AI contribution, leading to different reactions and assumptions towards the images and what the labels implied about them. And lastly, Theme 4 describes what were the participants' motivations to disclose and what influenced them.

Theme 1:
The capacity of ideation is the core of Human Authorship

Across interviews, participants consistently described their value as content creators as residing in their role as ideators: the originators of the concept and intent behind their social media content. For participants, this act of forming the concept defines what they understand as 'authorship'. In this sense, ideation is what defines the boundary creators use to distinguish between legitimate human contribution and AI, since ideation secures social recognition and value under perceived displacement.

"Human work is like gourmet food and AI work is like fast food".

Participants also described human ideation as valuable because it is tied to effort and craft, often using metaphors to express what AI cannot provide. For example participant #1 compared human made work to food that "nourishes" because it contains visible effort and intention: "it nourishes more… You know what it contains. It tastes richer". Similarly, Participant #3 compared "fast food vs gourmet" to describe why human craftsmanship may be revalorized over time: "Fast food exists, but gourmet food exists as well".

For some participants, AI was generally acceptable as a tool when it supported execution, refinement, or visualization with the condition that the creator remained the source of intent. This baseline frames that authorship is attributed to whoever originates the meaning and intent behind their work. Participants evaluated both AI tools and disclosure labels through that boundary (Theme 2).

Theme 2:
Labels like "AI-Generated" and "Made with AI" communicate full AI-contribution

During the think-aloud sessions, creators did not perceive labels as mere descriptions; instead, the labels functioned as cognitive cues that reshaped how the images were interpreted. The labels 'AI-generated' and 'Made with AI' offered a clear account of AI-involvement, effectively setting an initial hypothesis about the image being either generated with AI or made with AI. Consequently, participants scanned the images thoroughly for synthetic cues (e.g., lighting inconsistencies, anatomical distortions, or texture smoothness); however, they found organic cues indicating human-provenance instead. Regardless of this tension, participants noted that AI-generated imagery has become increasingly realistic, and ultimately decided to validate the label's claim and treat it as a reliable warrant.

"AI-Generated"

When experts encountered the "AI-generated" label, it acted as a signal that the content was fully synthetic. Interestingly, the label triggered professional skepticism that overrode their own visual expertise. Even when a creator's eyes told them an image was a realistic photograph, the label forced them to ignore their intuition and search for AI cues.

> *"If I didn't have the label, I wouldn't think it's AI at all… it does look quite realistic." (Participant #3)*

> *"The organic touch this image has makes it look more realistic than other AI images. Without the label, for me it could be a photo from a political meeting, right? So yeah, the label helps me." (Participant #1)*

"Made with AI"

The "Made with AI" label was often normalized and interpreted as a specific descriptor. Instead of prompting participants to wonder who was the author, they immediately inferred with clarity that the image was synthetic because the label implied it:

> *"This tells me that everything was made with AI. Like, basically, one hundred percent, one hundred percent AI." (Participant #2)*

Similarly, the label "Made with AI" was understood as "AI was part of the workflow," without threatening participant's notion of authorship. Another observation was that the genre of the image can contribute to diverse judgements about authorship. In this context, AI was mostly accepted as the tool to produce the image; to aid in visualization or execution, while the underlying idea and intention remained human. This explains that the genre can similarly play a role in what the participant considers a threat to authorship.

> *"A render made with AI by an architect... seeing this label... it feels normal." (Participant #3)*

In summary, precise labels reduce ambiguity and allow creators to confirm the image's origin through visual evidence rather than speculation.

Theme 3:
Labels like "Co-created with AI" and "AI-assisted" do not clarify AI's role, inviting speculation.

Labels such as "Co-created with AI" and "AI-assisted" are often insufficiently specific to communicate the AI's contribution with precision. This can lead users to form assumptions about the role played by AI and the human. In fact, when participants observed the labels, they reacted by creating a story about who was the author (the human or the AI). Consequently, they probed the images for human vs. AI contributions to resolve questions of origin, credit, and authenticity.

While "AI-generated" and "Made with AI" triggered strong origin hypotheses, "Co-created" and "AI-assisted" were less specific and led participants to infer their own assumptions regarding who deserves the credit of the image: the AI or the human? This effect was especially visible in artistic contexts, where authenticity and creative intent are primary sources of value.

"Co-created with AI"

Participants understood this label as 'shared contribution', but simoultaneously, they found confusing the precise extent of AI contribution and how

In artistic contexts, the "Co-created with AI" label triggered a need for clarity in the extent of AI contribution to understand "who did what", since the label did not provide details about the extent of AI contribution the same way 'AI-generated did', participants often inferred their own assumptions.

> *"What was made by humans and what was made by AI?" (Participant #3)*
>
> *"And co-created could mean that you made something, the AI made something else, and then you took what the AI produced and adjusted the colors" (Participant #2)*

"AI-Assisted"

This label frames the image as human-authored with AI assistance, but does not imply how (a human used AI in some way). This label leads to different interpretations. While some participants assumed that a human used AI to transform real material, it often led to discomfort.

> "I think that as human creators we have the ability to make images like this or better. I think that's the part that affects me the most as a creator." (Participant #1)

Theme 4:
The act of disclosure is meaningful in theory but feels like a risky reputational negotiation in practice

This theme aims to answer SQ2: Which factors influence a creator's willingness to disclose AI involvement on social media? The findings indicate that a) what creators found meaningful about AI disclosure and b) what they found risky indicates what motivates them to disclose and what is still a gap.

Participants value transparency and protecting audiences

Participants described AI disclosure as a practice aligned with personal integrity and transparency. For many, labeling AI use was perceived as the "honest" choice, particularly because social media audiences are seen as vulnerable to misinformation or misinterpretation when synthetic content is presented without context. In this sense, disclosure was understood as a form of responsibility to protect viewers from being misled and an act of transparency about their intent.

Yet, disclosure is a risky negotiation of reputation and security

Some participants feared that disclosing any AI involvement could lead to an exaggerated attribution of AI agency, leading audiences to assume the creator "did everything with AI," even when the human contribution (especially ideation and intent) was central.

Additionally, creators described disclosure as occurring within a platform context where non-disclosure may carry penalties, such as reduced reach, content removal, or account sanctions. This outlines the pattern that the act of disclosure is not only experienced as a voluntary practice, but as a strategic response to platform enforcement.

> *"And sometimes Instagram already tells you: 'No, this contains AI,' ? And it makes me feel like, now people are going to think that almost the whole photo is AI,' you know?" (Participant #4)*

7.2 Cross-cutting theme dynamics

This section synthesizes the connections between themes by answering the research questions that motivated the study. The insights suggest that labels shaped how creators attributed authorship, how they evaluated the image and what they considered were the risks and benefits of disclosure.

SQ1: *How do creators interpret what AI disclosure labels imply about who made the image (authorship and credit)?*

Since the study employed four different labels, the findings revealed that they were interpreted differently due to their framing variations.

a) Precise framings ("AI-generated," "Made with AI"; Insight 2) function as dominant codes (Hall, 1973), confirming provenance by overriding visual cues and affirming AI

contribution and overriding perceived human effort or intent.

b) Labels with imprecise AI framing (Insight 3) act as unresolved messages that raise questions about who authored the idea: Labels such as “Co-created” and “AI-assisted” do not specify the extent or role of AI contribution and led participants to look for missing details on the image and arrive to their own conclusions about who was the author or had the credit of the image.

SQ2: *Which factors influence a creator’s willingness to disclose AI involvement on social media?*

The findings highlight that participants understood current label practices as both an act of transparency and a risky practice. This contradiction stems from the fact that the complexity of the participants’ workflows does not currently match the options that certain social media platforms provide to disclose AI use, which may indirectly lead to reputational/social risks and sanctions from the platform. However, participants argued that they value transparency, therefore they weigh the act of disclosing against reputational and platform risks. This reveals that the content creator is in a constant negotiation about rule compliance and the ethical need to be truthful towards their audience and their work.

AI disclosure may seem motivating when it is viewed as an ethical and necessary practice to avoid misleading audiences (“otherwise it’s fraud”) and aligning with their authenticity goals. At the same time, willingness to disclose may decrease when creators anticipate stigma or misattribution, particularly when disclosure could be interpreted as “AI did everything,” regardless of the complexity of their workflows. In addition, creators described platform enforcement as a structural factor in which non-disclosure may carry penalties.

In summary, willingness to disclose is a negotiation of ethical transparency, authorship protection against anticipated consequences (audience judgement and platform governance).

7.3 Chapter Summary

Summary of the four themes that emerged from the thematic analysis:

- Theme 1: Authorship = Ideation. Creators defend the “Idea” as the human element.
- Theme 2: Precise Labels (‘AI-Generated’ and ‘Made with ‘AI’) act as dominant cues that confirm synthetic origin.

- Theme 3: Imprecise Labels (‘AI-Assisted’, ‘Co-Created with AI’) act as ambiguous cues that lead to confusion about credit.
- Theme 4: Disclosure is theoretically meaningful (transparency) but practically risky (platform sanctions/social stigma).

DESIGN SYNTHESIS

PART IV

8 DESIGN SYNTHESIS

This synthesis offers a synthesis of the literature review and the findings. It begins by a reflection about labels being the lens through which creators' intentions, audiences' interpretations, platform governance, and regulatory demands overlap.

8.1 Scope and lens

The practice of labeling AI-generated content involves content creators, users, the platform, and the constraints of AI regulations. The EU AI Act mandates social media platforms to implement AI disclosure labels with the aim of enforcing transparency about the use of AI by informing audiences when content has synthetic provenance (European Union, 2024, Art. 50(4)). However, platforms are socio-technical ecosystems in which human and non-human actors co-produce meaning through negotiated interactions in a reputational economy. As a result, platforms mediate how transparency, authorship and creativity are understood in an environment where AI-generated content proliferates at scale (Van Dijck, 2013, p. 12; French, 2025; European Parliament, 2025).

Therefore, AI-labels are technical features that do not simply act as disclosure mechanisms, but transcend how the meaning of content is communicated to users. From Stuart Hall (1978)'s analysis in semiotics, this thesis understands AI-labels as signs that communicate both a denotative (a literal claim) and connotative (a subjective claim) claim about the level of AI involvement (AI or no AI), who authored the work and how much human effort it contains.

In this regard, language representation through signs (like AI labels) enable others to read, decode, or interpret the message of the content (Stuart Hall, 1997). This approach implies that meaning is produced and disttributed across the social media ecosystem, and thus the AI label becomes a constitutive element in the communication process that shapes social subjects and how users understand/interpret multimedia content. As demonstrated, research from Gamage et al. (2025) found that, among the various dimensions within the design space of a label, its wording can act as a semiotic sign in the social media feed that frames how content is perceived. Consequently, this implies a potential gap or mismatch between what the creator intends to communicate with their content and how AI labels alter the intention of this message. Therefore, such inferences operate as judgements of authorship in the feed, that shape how credit, effort, and legitimacy are attributed to the creator(Hall, 1978; Burrus et al., 2024; Rae, 2024; He et al., 2025).

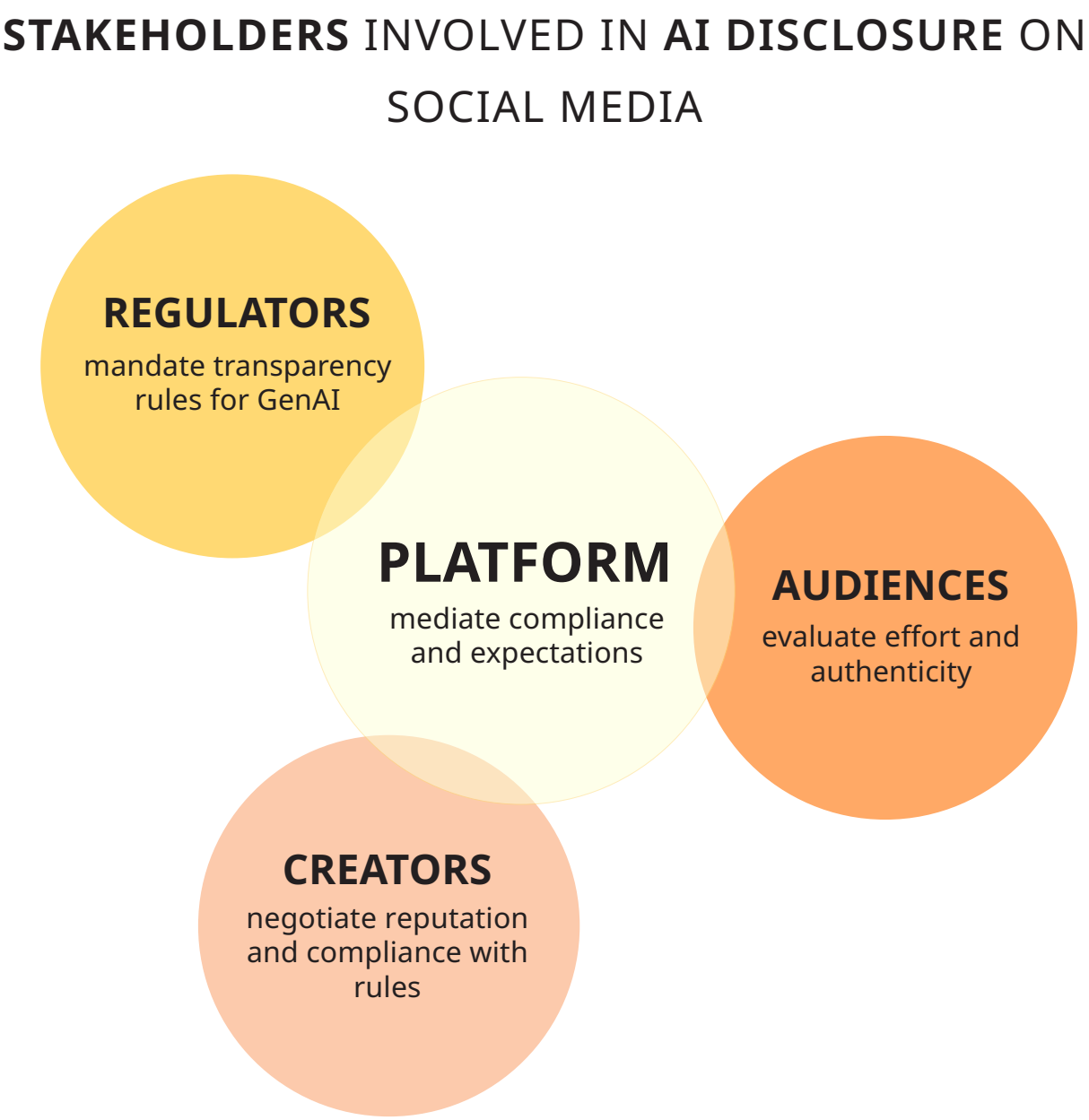


FIGURE 8.1. Overview of key stakeholders in AI disclosure on social media. The platform represents the overlap between regulatory transparency goals, creators' disclosure considerations, and audiences' evaluations of authenticity and effort.

Figure 8.1 illustrates how regulators initiate the circuit by defining transparency norms and obligations for AI-generated content. Platforms mediate these by implementing disclosure mechanisms (such as specific wording, icons, and interaction designs) and integrating them into ranking, moderation, and compliance processes. Creators respond by incorporating these mediated disclosures into their practices, balancing them against concerns for reputation and legitimacy; the AI label thus shapes how authorship is enacted and perceived in content feeds. Audiences, in turn, decode the label as a semiotic signal, assessing authenticity, effort, and trustworthiness in ways that may differ from creators' intentions. These inter-

pretations produce feedback via engagement metrics, comments, complaints, and evolving norms, which in turn prompt adjustments by platforms and refinements in creator strategies. Through this circuit, AI labels emerge as a shared mechanism for meaning production across regulation, platform operations, creator identity, and audience reception.

8.2 What are the constraints for content creators in the social media ecosystem?

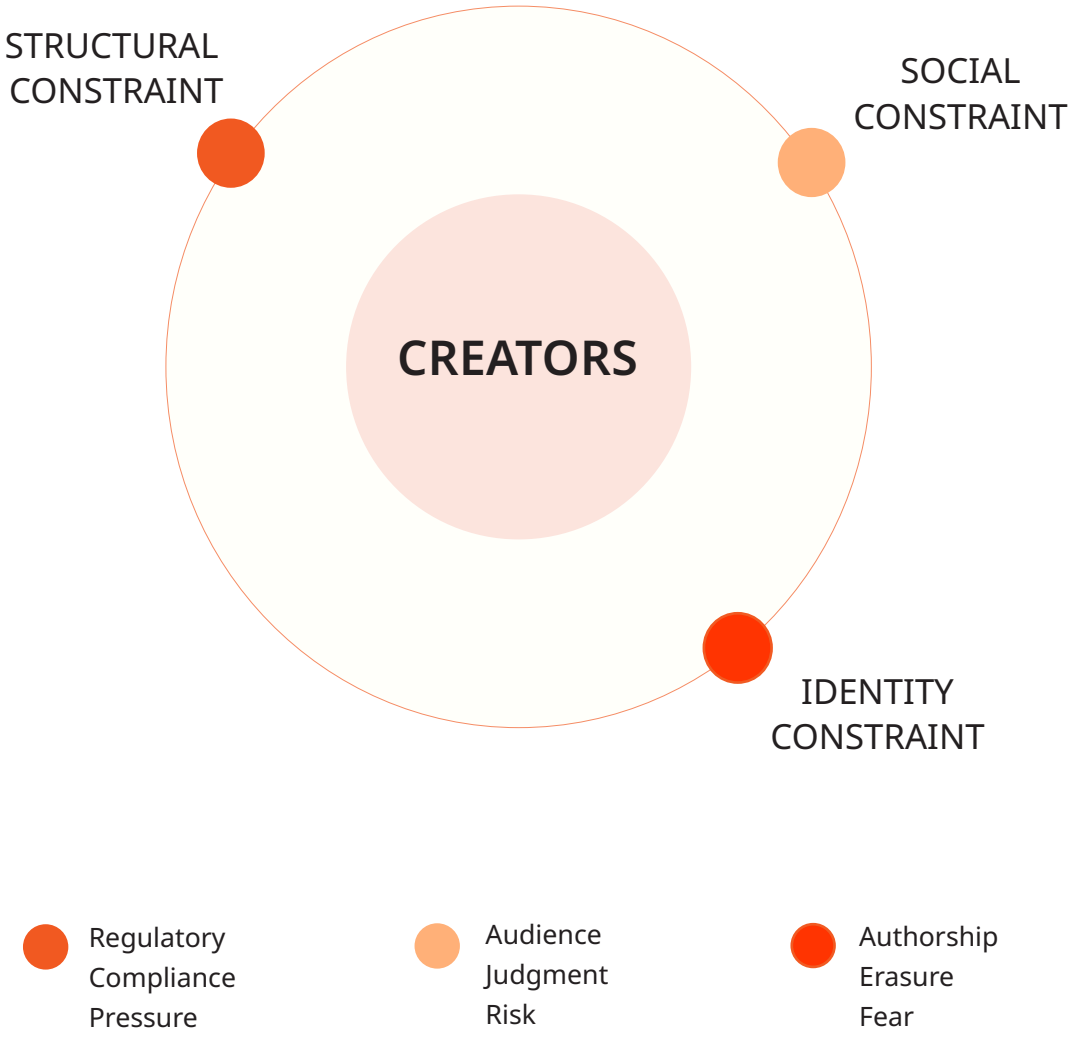


Figure 8.2. Creator constraints diagram influenced by structural constrains, social constraints, and identity constraints.

Content creators operate inside a platform environment where visibility, meaning, and compliance are largely shaped by platform rules and interface protocols, while also being pres-

8.3 How do labels challenge the concept of authorship and creativity?

A ‘message’ refers to the unit of communication that conveys meaningful information (Duck & McMahan, 2020). Social media platforms serve as a medium in which content creators communicate to their audiences through content (e.g. images, videos). However, while this communication form acts as the primary visual sign, it is surrounded by secondary signs that frame how that content should be read. AI labels are an example of such a signifier:

Signs can frame the meaning of the message that is being communicated in two ways: denotative and connotative. The denotative level refers to the literal, explicit meaning (technical fact: ‘this content is made with AI’); and the connotative level encompasses the psychological, emotional, and ideological meaning that a sign evokes when interpreted by the viewers (‘the entire content is made with AI’). Therefore, an AI label can consolidate a denotative claim about the content’s provenance, and simultaneously, a connotative claim that inspires diverse understandings and assumptions about the content and the creator who posted it (Stuart Hall, 1973). Consequently, there is a mismatch regarding what the creator intends to communicate and what the AI labels imply about their content. In fact, AI labels can act as dominant signals that frame AI involvement, guiding user’s attention strongly towards the presence of AI (Rae, 2024).

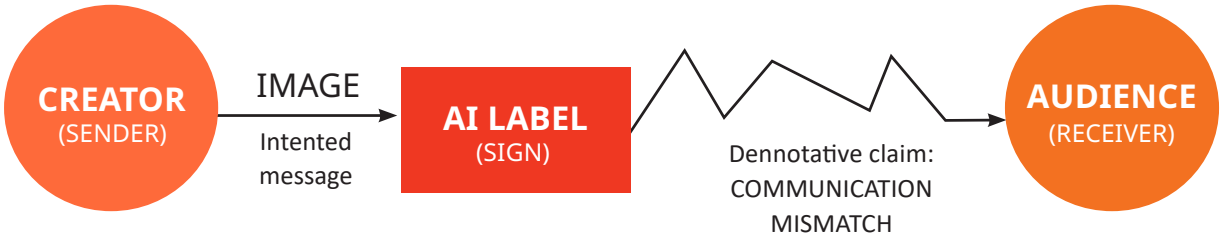


Figure 8.3.a illustrates how an AI disclosure label functions as a sign in the feed: it denotes a literal provenance claim (AI involvement) while also triggering connotative inferences (e.g., ‘fully AI,’ ‘low effort,’ or ‘less credible’), which can reorient the audience’s reading away from the creator’s intended message, producing a mismatch (adapted from Hall, 1973).

Alternatively, as found through the Design Study of this thesis (Theme 1) reveals that creators anchor their sense of authorship in ideation and intent. For creators, the act of forming the concept and defining the intent is the non-negotiable boundary between human contribution and AI. Labels challenge this notion by centering the focus on AI.

Figure 8.3b. Visual representation of "AI-Generated" as a dominant sign that overrides human cues (e.g., participants scanning for synthetic errors but trusting the label instead).

In contrast, **labels like “AI-assisted” or “Co-created with AI”** functioned as ambiguous cues, prompting participants into an active process of "negotiating meaning," where they had to guess who authored the content. For example, "Co-created" caused friction because it implied an "equal partnership" but did not explain how or to which extent, which many creators felt overvalued the machine's role in a human-led workflow.

Figure 8.3c. Visual representation of "AI-Assisted" as a sign that prompts the viewer to question who created the content.

In this thesis the findings showed this dynamic: labels such as “AI-generated” and “Made with AI” led to strong provenance interpretations. Participants treated these labels as dominant signals because they were explicit and categorical. These signals influenced how participants assessed the image, including in cases where the visual cues looked organic.

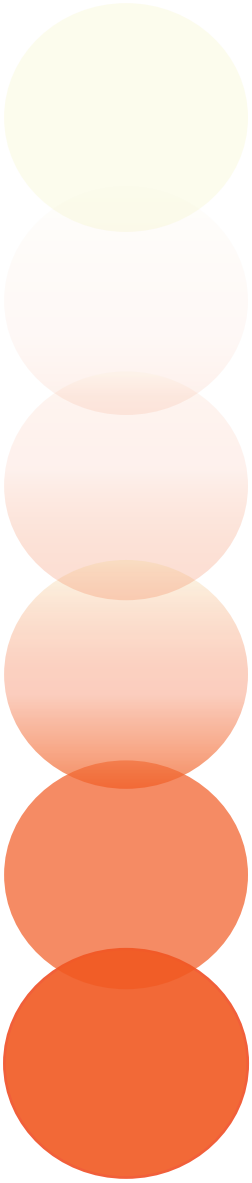
In contrast, imprecise labels were interpreted through what Hall (1973) describes as negotiated codes. When AI involvement was presented as partial or unclear, participants relied on their own interpretations: inferring the relative roles of human and AI and adjusting authorship judgments by genre and context. Similarly, framing AI next to the image will either way let users conclude that AI was involved somehow, and this is usually tied to concepts of AI-generation or AI-ideation, therefore, lack of effort.

8.4 How complex is it to label AI generated images?

Labelling AI involvement in social media content is complex for two reasons: (1) AI involvement is not binary, and (2) AI involvement is distributed across stages of a workflow.

The Spectrum of Automation: From Manual to Synthetic

AI-involvement is not homogenous and involves a spectrum that stretches from fully manual, AI-assisted, Co-created to Generated with AI and fully synthetic. Parasuraman, Sheridan & Wickens (2000), argue that automation comprehends a continuum of levels of automation interacting with human operators, from low to high; the first level indicates full human intervention, and the tenth level equates to full automation. In the case of images, the levels would range from natural to fully synthetic as shown in Figure 8.4a.



HUMAN-MADE (no AI)

Natural photo/illustration/collage, including standard digital editing (crop, retouch, color grade, dodge/burn).

HUMAN-MADE/ AI-ASSISTED EDITING

Denoise/upscale, generative fill, remove object

HUMAN-LED/ AI-ASSISTED CREATION

AI used for moodboards/variants/reference;
human redraws, reshoots, models, or heavily reconstructs.

CO-CREATED (Human and AI contribute equally)

Text-to-image / image-to-image; human constraints and decides, iterative prompting

AI-GENERATED (Human mainly curates/ publishes)

One/few prompts → human selects best output → minor refinements (Human edits)

FULLY SYNTHETIC (Minimal, no human direction**)**

AI-Generates many images at once, template prompting, production at scale.

FIGURE 8.4a. The Spectrum of AI Involvement (author's synthesis). Conceptual model developed in this thesis, informed by levels-of-automation work (Parasuraman et al., 2000; Sheridan & colleagues) and human–AI co-creativity / mixed-initiative frameworks (Yannakakis et al., 2014; Rezwana et al., 2022).

Furthermore, the cognitive interaction between human and machine occurs in distinct phases, Parasuraman, Sheridan & Wickers (2000), name four parallel cognitive stages that occur in the interaction between the human and the machine in the process of automation, such as information acquisition, decision and action selection and action implementation. Rezwana & Maher (2022) parse these concepts into the phases of human collaboration with AI in media such as:

Stage 1

Direction and Constraint Setting:

Setting the intent by defining constraints, prompting and defining character attributes and style rules before or alongside making outputs.

Stage 2

Generation:

The AI generates outputs after the conceptual direction. This is often refered to the "Black Box".

Stage 3

Selection/Evaluation/ Refinement:

Evaluation can be done by the AI (as evaluator) and it commonly leads into refinement (polishing).

Case Study: The Iterative Workflow of Chiara Di Lodovico.

As illustrated by Design Researcher Chiara Di Lodovico, generating content with AI (image to image or text-to-image) is a high-effort, iterative process where human agency fluctuates across three distinct stages:

Stage 1. Direction and Constraint Setting (High Human Agency).

Control over the process is exerted through specific parameters such as implementation of keywords (e.g. design, minimalist), controlling parameters such as the weight, using mood boards to enforce a visual language; uploading a set of images with similar aesthetics or asking the system to describe the image in order to generate the prompt. Di Lodovico explains this stage in two phases:

On parameters:

“Midjourney has a lot of controls, so you can also include a character and you can decide the weight of the character into the picture. Or you can now have a new feature which is mood

boards. So you can upload one image or a set of images that speak the same visual language, and then you can create images with that specific language by adjusting the weight

the moodboard has. So for me, Midjourney is the thing that I can control the most in terms of parameters."

Similarly, this stage often involves a recursive process where the creator translates intent into machine-readable prompts.

On keyword input:

"I extracted some keywords and I tried to put the keywords into MidJourney and at a certain point I found an image that made me think of an image I already saw. So I looked for it on Pinterest. I saw what kind of words were used[…] I couldn't use another user's image, so I asked the system to describe the image, which is something that this system can do to extract a prompt, and then I also played with different versions."

Stage 2. Generation: The 'Black Box' (Low Human Agency).

Di Lodovico describes this as a "limbo" where she must accept a lack of direct control over the machine's immediate output. At this stage, the user relinquishes direct agency to the AI's probabilistic engine.

On "the Limbo":

"I give up some of the picture sides because I know that it would become a nightmare to control some things […] I will not have full control over the image and I will just stay in that limbo."

Stage 3. Evaluation/Refinement/Selection (Hybrid Agency).

The final stage involves regaining control through iterative manipulation, such as "zooming" or "inpainting" to build a scene. The AI output is treated as a "base" that is then refined through human intervention, such as inpainting or "zooming out" to build a larger scenario.

On Refinement:

"I control the process, but less the output […] I was more manipulating images for creating a scenario. I was creating, first, the base object and then zooming out and adding/editing the

picture with two people and then zooming out and creating the outer."

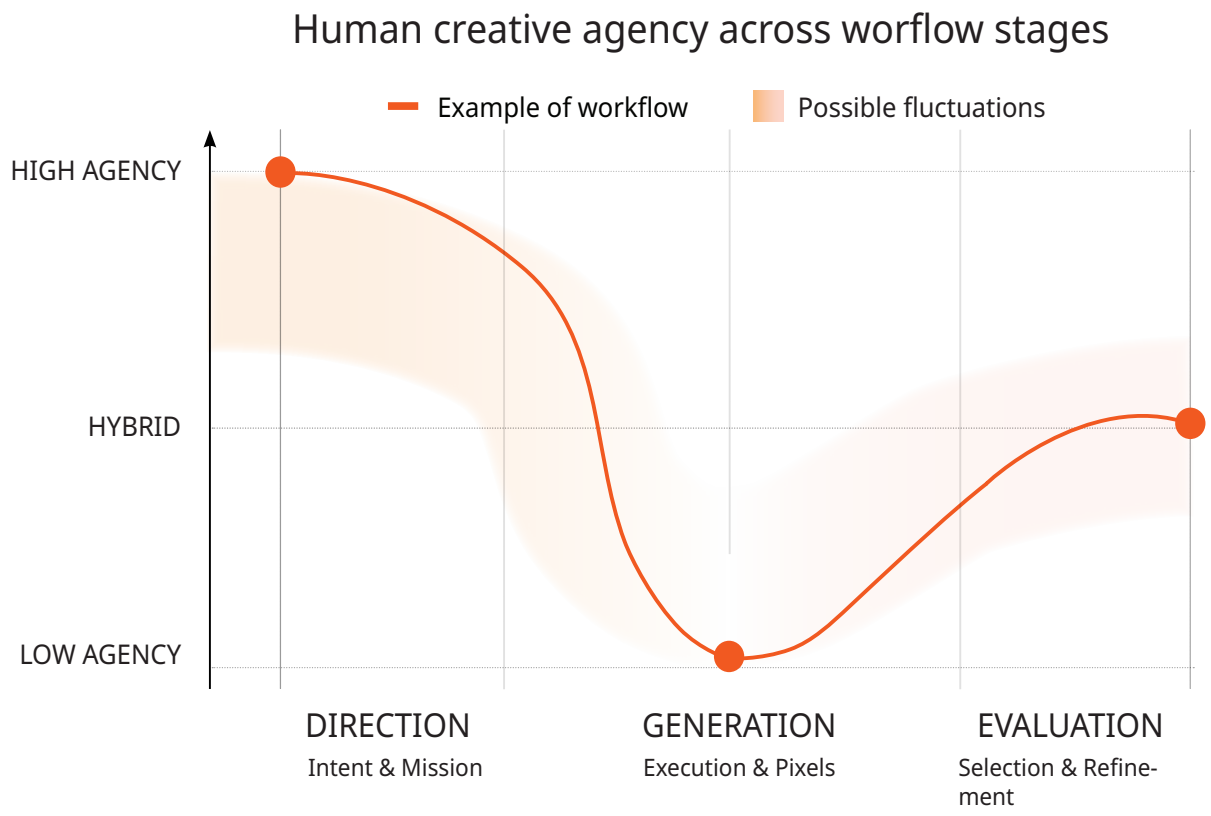


FIGURE 8.4b Human creative agency across AI image-generation workflow stages (author's synthesis). The curve illustrates Chiara Di Lodovico's shifts from high agency during direction and constraint setting, to low agency during model generation (the "black box"), and back to hybrid agency during evaluation and refinement. The shaded band indicates plausible variation across workflows and creators.

Figure 6.4b, based on Chiara Di Lodovico's interview, confirms Ananny & Crawford (2018)'s statement that accountability is a relational process in the context of automation. For instance, the Generation Stage is usually a "low human agency" process; Chiara describes this stage as a "limbo" where she accepts a temporary loss of control. In the Refinement/Evaluation stage, total agency over the output might not return, but a relational form of control occurs: through human actions such as selecting, steering, editing, zooming out, building scenes iteratively ("creating the base object... zooming out and adding/editing"), but Chiara mentions that she explicitly does not regain full control over the output.

Across workflows, agency is transferred and transformed. Therefore, accountability is a relational process because it depends on the fluctuating interactions and negotiations between human intent and algorithmic output across the entire workflow, rather than being a static property hidden inside the system itself.

On this account, the practice of AI disclosure labels underscores the complexity of visual content and intends to summarize it such as "Made with AI" and "Generated with AI" by explaining uniquely the generative phase. This formulation, construction of language or 'wording', of current AI labels prompts users to emphasize their attention on AI. As Burrus et al. (2024) state, the design problem of AI labels is not merely whether to disclose AI use, but how to communicate degrees and moments of intervention without overloading viewers' mental models and by helping them understand the accurate meaning of the image they represent.

6.5 The Paradox of Transparency: Illusion vs. Truth

Defining Transparency in the Era of AI

The concept of transparency is linked to an ideal of absolute disclosure, but in practice it is not a "precise end state" in which everything is visible to the user. More precisely, t**ransparency becomes a selective form of observation of the stage of a process** (Ananny & Crawford, 2018). Stohl et al. (2016, as cited in Ananny & Crawford, 2018) argue that one of the limitations of implementing transparency (AI disclosure) on digital systems is that the information displayed on the screen will unintendedly cause that unimportant details become the focal point of the user's attention, consequently, creating a distraction that conceals the more critical information regarding the creator's actual intent and labor.

Therefore, within platform interfaces, the ideal of transparency collides with the logic of immediacy in interface design, where digital elements aim to disappear into the experience, so interaction feels natural and intuitive, while meaning is produced through a "play of signs" that users interpret in divergent ways (Bolter & Grusin, 1998). AI disclosure labels, thus, intervene in the user's state of flow by reintroducing mediation at the moment of viewing the content; this is called hypermediacy.

Deriving from Ananny & Crawford (2018)'s research, transparency can be bifurcated into technical transparency or aesthetic transparency. AI labels, therefore, enact the version of technical transparency to indicate provenance; conversely, social media content (images) enables aesthetical transparency, where the experience fades into an illusion of immediacy, such as the experience of real physical objects.

The paradox of transparency

In the context of AI labels, stems from regulatory bodies demanding clarity in a digital space where audiences infer intent, quality, and authorship from minimal cues. In this

regard, ambiguity is a design resource that supports assumptions and interpretation rather than providing certainty (Rae, 2024; Gaver et al., 2003). When technical transparency is forced upon an aesthetically transparent image, it triggers a ambiguity about authenticity, prompting users to look past the label to look after truthful data in the image.

Images as engineered illusions

> *"To achieve photorealism, the synthetic digital image adopts the criteria of the photograph" (Bolter & Grusin, 1998).*

Throughout history, visual media has embedded the concept of immediacy, by creating the illusion that digital representations are as real as the physical world. Techniques such as linear perspective, photographic optics, and later digital images, have engineered conventions that minimise awareness of mediation in order to produce a sense of presence. In this regard, digital images are designed to look real through edition and retouching, depth of field, and culturally learned expectations about what realism should look like (Bolter & Grusin, 1998).

Realism ≠ Authenticity: Why Visual Judgement is Unreliable

This enactment of technical transparency through AI-labels prompts users to divergent assumptions. During the think-aloud session of the Design Study, participants, as skilled viewers, responded to labels by engaging in visual scrutiny of the images, while seeking after specific heuristics:

Anatomical Anomalies (hands):

A pattern emerged during the think alound session in which anatomical coherence became a marker of authenticity. One of the participants explicitly mentioned looking for anatomical errors and scrutinized the fingers, noting that one hand more fingers than it should: *"I'm a bit… overanalyzing it, I'd say. Like, no artist would draw seven fingers on an angel, right?"*

Light and Shadow Inconsistencies:

The lack of light and shadow coherence serves as a critical heuristic for viewers evaluating the physical validity of a scene. Consequently, participant #2 considered the structure and lighting of wings as a diagnostic tool, identifying inconsistencies in how they are rendered compared to natural form: *"The wings that are turning toward the front aren't being hit by the*

light — I mean, they're dark — but the back wing, which has the same angle, should be getting the same light. That one is being hit by light, and the others are getting shadows. "

Perfection as a synthetic cue vs Organic forms as authentic cues:

Participants shared similar beliefs about what AI imagery looks like regarding what they considerer 'organic' versus 'too perfect'; they argued that excessive perfection such as flawless symmetry, unrealistic smoothness indicated synthetic origin. As one participant explained: *"If it were Artificial Intelligence—a bit more, uh, 'normal'—well, it would all look perfect, like everything was very planned. So that's why I think that is the organic touch this image has, which makes it look more realistic than others."* Similarly, another participant argued "*if an image were AI, it would look perfect and very planned*".

Photographic conventions (camera focus, photographic blur, linear perspective)

In alignment with Bolter & Grusin's assertion that digital images are designed to look real through learned conventions like depth of field, participants used photographic "blur" to verify realism:

"The image looks quite realistic, especially because the focus is in the foreground and the background has that natural, casual blur—like any camera would do when it wants to give prominence, in this case, to that old man who's there, instead of what's in the background" (Participant #5).

However, contemporary AI models replicate these cues, therefore the viewer's scanning strategies become unreliable for inferring provenance. For instance, the first probe in the Design Study features a real image paired with the label 'AI-generated', proving that while AI disclosure can be necessary to correct provenance when visual evidence is insufficient, it can also prompt viewers to inspect the pixels for confirmation, displacing attention from intent and labor to superficial cues detection. Consequently, this reproduces the distraction paradigm described by Stohl et al. (2016, as cited in Ananny & Crawford, 2018) that transparency, as a tool for disclosure, can reallocate attention, often toward cues that feel diagnostic for human provenance yet are now simple to replicate by algorithms. Therefore, social media strategies for enacting transparency do not necessarily produce truth about provenance (Burrus et al., 2024; Rae, 2024; Gaver et al., 2003).

To further elaborate on this argument, the following collection of images, some of them real and some synthetic, explain this in practice:

REAL OR NOT?

Figure 8.5a: **Natural Photography with Minimalist Composition.** This photograph creates a 'synthetic illusion' by isolating the subject against a high-contrast environment where the loss of atmospheric depth make the cabin appear as a 3D asset placed on an infinite heaven.

Figure 8.5b **Synthetic Landscape via Midjourney.** This figure illustrates a mundane composition) with diffuse atmospheric lighting. The presence of unpolished details mimicks the aesthetic of a candid photograph.

Figure 8.5c: **Natural Photography.** This image, though a real photograph, demonstrates how modern mobile image processing can create a 'synthetic' aesthetic. The extreme saturation of the organic elements (tomatoes and arugula) and the high-frequency sharpening of the textures mimic the appearance of a digital render.

Figure 8.5d: **Synthetic Image via Midjourney.** This figure is on synthetic origin, yet the realism it evokes is derived from the complex light interaction within the refraction in the water glass, the source reflections in the wine and the organic texture fidelity of the food.

Figure 8.5e **Natural Image with High Digital Latency.** Although this is a captured photograph, it exhibits 'synthetic artifacts' due to heavy post-capture processing.

Figure 8.5f. **Synthetic Urban Environment.** This image is generated with Midjourney, however it achieves 'perceptual realism' by simulating specific photographic constraints.

REFLECTION

PART V

9 Takeaways

This section does not offer design guidelines, since AI disclosure practices, platform policies, and audience norms are continously evolving in the times of AI. Alternatively, this thesis translates the insights into transferable takeaways for peers, educators and a final reflection.

Messages to peers:

Message 1:

Disclosure labels as semiotic lenses: Mediating the appreciation of content

To design for AI disclosure in the social media ecosystem, focusing on the interactions among stakeholders leads to an understanding on how meaning is produced, distributed and understood (Van Dijck, 2013). Stuart Hall (1997) explains how meaning is a construction that results from the signs the medium provides. This medium is the interface, rich in signs that hold technical and cultural information. Bolter and Grusin (1999) understood this medium as the recipient in which media (e.g. images, videos or text) immerse the user into an experience that blurs the concept of what is considered 'real'. The illusion of immediacy is what contributes to meaningful interactions among users (both content creators , as producers; and audiences, as receivers).

Regulators and platforms introduce the rules that govern the way meaning is conveyed. AI-labels are implementations which often act as the interruptors of this immediacy by reminding the user that their experience might be, at times, inauthentic. Hall (1997) indirectly explains how the language that labels employ to represent content, such as "AI-generated", "Altered" or "Made with AI" can either reflect a truth about the actual provenance of media, or, conversely, produce a new meaning about it. This leads to the question, then, about what is the semantic meaning of the word 'AI' and is it an accurate representation about content? Moreover, what is the correct way to introduce AI labels, as hypermediacy, in the context of AI related content in social media?

As Burrus et al. (2024) report, many viewers don't know what "generative AI" means and ignore the complexity of the spectrum of AI-involvement. Wittenberg et al. (2025) attempt to solve this conundrum by introducing the concept of 'process labels', that reveal the process in which AI was involved. Attempts from industry standards, such as the Coalition for Content Provenance and Authenticity (C2PA), aim to introduce a 'less binary' approach by applying UX principles such as progressive disclosure, to reveal the complexity of content provenance. However, the conflict might remain if the first layer on the interface-level persistently incorporates the word 'AI' as a dominant preattentive cue. This prompts leads to the question, how to frame the language of disclosure so it reflects the complexity of the workflow without systematically erasing the creative identity of the creator?

Message 2:

Designing for radical transparency in AI labels might be counterproducent, but designing for ambiguity might be a resourceful strategy.

To which extent is AI disclosure appropriate considering the full spectrum of AI? While Aler Tubella et al.'s (2019) propose the "glass box" model to counter the "black box" (the opaque side of AI generative processes) in order to encase AI in verifiable input-output norms derived from moral values such as transparency, Ananny & Crawford (2028) note that this ideal does not reflect the reality of AI disclosure, since full disclosure can be a harmful practice in the sense that it might not necessarily lead to more understanding and thus can distract user's attention. On one hand, Gamage et al. (2025) found that labels that describe processes can cause cognitive overload on users, and on the other hand, Burrus et al., (2024) emphasized that labels can overcenter the attention strongly on AI. Such collapse can undermine perceived credit and effort in a social media ecosystem where AI generated media is challenging the notions of copyright, ownership and authenticity (French 2025; Rae, 2024; Burrus et al., 2024). Similarly, during the interview, Chiara Di Lodocivo noted that disclosing the amount of prompts to prove effort and intent can also lead to what she considers 'The Sustainability Paradox', considering that producing AI-generated content is environmentally costly, inviting negative audience interpretations if this was disclosed. Therefore, this leads to the question:

Which levels of AI-involvement merit visualisation via interface labels?

He et al. (2025) proposed that disclosure should distinguish between the nature of the contribution (whether the AI provided the substance of the ideas or merely the stylistic polishing). Currently, AI label practices emphasize on the stage of generation; with the exception of some social media platforms such as TikTok, which already parse the nuance between 'AI-assis-

tance edition' and 'AI-generation'. This results in new tags specifically designed for filters that are discretely segregated from 'AI-generated' labels. Accordingly, this can be operationalized as presenting diverse levels of AI-involvement as separate elements within the interface, rather than collapsing the totality into 'one-size-fits all approach'. In Gestalt principles, this could equate to separating the spectrum into different visual clusters. Another approach, as proposed by Gamage et al. (2025) is contextual disclosure, depending on the content-type (e.g., news, humorous content, art). Hence, the implementation of full disclosure into one single label could be avoided, and designers could play with ambiguity in a favorable way by creating the illusion that the levels of AI-involvement are not a gradient, but distinctive categories. In fact, Gaver et al. (2003) argue that ambiguity can actually be a favorable a resource for UI design if implemented deliberately.

Message 3:

From traditional craft to automation: notions of authorship have always been evolving

French (2025) posits that throughout history, the notions of authorship and creativity have never been static and are inherently fluid. Manovich & Arielli (2024, p.32-37) observe that within contemporary social media ecosystems, reputational credit is negotiated through a complex interplay of ideological frameworks, social media metrics and what is considered meaningful to users. Consequently, a critical tension emerges regarding the methods used in the production of content.

Nevertheless, the diversity of tools that facilitate creative expression have always been present. And although they bifurcate in methodology, they converge in the shared imperative of seeking the immersion of the viewer into the illusion of a real experience, irrespective of the production process.

For instance, traditional painting encapsulates a different process compared to the one photography for generating an output. Bolter and Grusin (1999) elucidate this: with traditional painting the author employs brushes to paint on a canvas, and with photography, the author must arrange the scene and press one button. Applying Bolter and Grusin's (1999) theory of remediation in the AI era, the same shift occurs: the author might use AI tools to enhance or refine content they produced themselves or iteratively interrogate a specific set of keywords to commission the AI to generate an image. However, the difference regarding AI as a method or tool, is that creative agency is more constrained and can be hybridized as a co-partnership between the AI and the human, which further complicates notions of authorship.

Much like photography was once critiqued for reducing the labor of art to the single press of a button, AI now exacerbates previous notions of labor by enabling assistance in every stage of the process described in this thesis, following the phases of human-AI collaboration (conceptualization, generation, evaluation/refinement) defined by Rezwana & Maher (2022). This creates a wicked problem for the contemporary author. He et al (2025) found that users value human intent and effort, so the more AI contributes to the production of content, the less credit the human author invariably receives.

Message 4:

Generating Images with AI carries two levels of responsibility: For the inputs provided to the machine and the social implications of the output

Generating AI images introduces a double layer of responsibility: first, responsibility for the inputs and production conditions (the data introduced), and second, responsibility for the social life of the output once it circulates. On one hand, Chiara Di Lodovico described a "sustainability paradox", explaining how the creative process often requires many iterations (prompts, variations, edits), but this increases environmental cost and can even make "showing the process" itself an unsustainable practice at scale. She also emphasized privacy and consent as central responsibilities, arguing that designers should avoid using identifiable people's images without consent, consider GDPR-sensitive data implications, and recognize that consent may not cover downstream transformations (e.g., creating realistic "replicas" or digital twins that go beyond the original intent). In fact, the General Data Protection Regulation (GDPR, 2016) on sensitive data emphasizes that personal data regarding natural persons is subject to stringer requirements (General Data Protection Regulation [GDPR], 2016, Art. 4).

On the other hand, Di Lodovico also noted how realistic synthetic media can lead to deception and manipulation in social platforms (e.g., fake profiles, staged "news"), meaning responsibility also includes anticipating how outputs may be interpreted, misused, especially when audiences dismiss labels or captions.

Messages to educators

Message 5:

Cultivating AI Literacy through relational human agency across the Spectrum of AI-Involvement

Defining the boundaries that situate AI as a tool rather than an author requires a nuanced understanding of the spectrum of AI involvement and its implications on human agency as a relational and fluctuating role. It is necessary, thus, to move beyond the dualistic notion and acknowledge that creative production exists on a spectrum, or continuum, where human intent and algorithmic execution coexist in varying degrees (He et al., 2025; Bomba and De Angeli., 2025*).*

Based on the think-aloud interviews, the session with Chiara di Lodovico and the literature review, the findings demonstrate that the granularity of AI involvement clarifies the extent to which AI functions as a tool and when it becomes an ideator. During the interviews, participants perceived AI as an auxiliar and based their identities as content creators on their ability to ideate; consequently, maintaining agency over the conceptual phase, which was seen as primary pillar of creative legitimacy.

Framing AI involvement as a continuum, can more clearly define the boundaries that situate AI as an assistive software rather than an author. This distinction is essential to protecting the integrity of human ideation and fostering skill development.

Nowadays, AI is a software, an assistant and a productor. Therefore, an instrumental attitude towards AI is crucial and could ensure that human agency remains most important phase : conceptual phase where ideation, direction and intent act as core pillar. This premise could similarly apply to the higher levels of automation in which the "black box" is an unavoidable situation during the generative phase where agency is paused or reduced; in this phase, the legitimacy of the work is not derived from the manual execution, but from the conceptual or evaluation phases. Manovich & Arielli (2021) argued that even when AI generates the content, the human still makes use of their 'aesthetical sensitivity' to influence the outcome. However, it is important to mention that in higher levels of automation, authorship, ownership and attribution cannot be viewed through the same lenses through which human made or AI-assisted content are viewed, thus require a more relational criteria. Therefore, defining when is the appropiate context to employ higher levels of automation is an effective way of ensuring ethical boundaries while teaching how to employ the tool skillfuly and with aware-

ness.

10 Reflection

Incorporating Bolter and Grusin (1999)'s analogy about traditional art and photography as aesthetical mediums for creating illusions, I would like to introduce the perspective of Manovich & Arielli (2024), which complemented this view in great detail by arguing that these mediums, traditional art, photography and synthetic images, have an effect on human perception since, in one way or another, influence our pattern recognition pathways. As Ware (2003) mentioned in his book Information Visualization: Perception for Design, these pattern recognition pathways occur in three stages that are set into motion one visual cues enter perception.

If the symmetric, harmonious features of an image offer more possibilities of immersion for the user, then what users consider 'organic' cues showing human provenance become become more relevamt. Manovich & Arielli (2024) then explain that every image, regardless of the medium through which it was produced, is a representation of culture, and every user participates in this. Therefore, in social media settings, imagery functions as content and is introduced by platforms under the concept of “life stories”, following the principle of connectivity, as van Dijck (2013) elucidates. However, when AI has already reminded users that their digital reality might be even more illusory than real, the ideal of ‘social connectedness’ is challenged. Furthermore, when AI-labels attempt to break the illusion of immersion in order to inform users that the visual medium is composed of synthetic provenance - defined as hypermediacy by Bolter and Grusin (1999) - , trust risks erosion and the culture of connectedness that social media ecosystems built is thus interrupted.

Consequently, the current structures through which content creators communicate cultural meaning are no longer communicating their intended meaning when AI-labels display technical data about AI; the encoder is no longer sending the intended message to the decoder when the structures of power encase the message in a form that does not match the original purpose (Hall, 1997). Therefore, the aesthetic value of social media imagery is becoming secondary to its technical provenance. This pivots the audience’s position from belief to interrogation, where they are no longer decoding the “life stories” from a post in social media, but scrutinizing a digital artifact. The interaction designer’s challenge is no longer just to facilitate immersion, but the question to designers is: how to provide truth about images without disrupting the user’s state of flow?

What occurs within the social ecosystem of the interface is challenging traditional methods of designing for users, necessitating a new approach where design enables a more human-centric implementation of AI (Verganti et al., 2020). Moreover, the social notions of authenticity, ownership and transparency no longer stand for what they claim and the new reality that artificial intelligence has introduced invites a new approach for desining interfaces. Ananny & Crawford (2018) advice against full dislocure, since it might not address more understanding, then designers should consider other methods for addressing this matter; a suggestion could be implementing the principle of amiguity deliberately, as Gaver et al. (2003) suggest.

I learned through this thesis that designing for every user in social media is almost an impossible task; as Hall (1997) sustained, receivers of content hold divergent and complex mental models. Gamage et al. (2025) later validated this by simulating a social media feed to analyze users perceptions on AI-labels within a controled study. Verganti et al. (2020) elaborated on how algorithms are now helping designers solve users's needs at scale by personalizing content for each one of them based on their individual data. Therefore, this raises the question of whether data informed design could be considered for designing AI labels (or other solutions) that address regulators' goals for transparency while balancing audience and content creator needs.

11 CONCLUSION

This thesis studies how the linguistic framing of AI disclosure labels on digital platforms influences content creators' perceptions of authorship and their motivations to disclose AI usage.

To address this, a qualitative methodology was employed while considering previous literature review. The core of the study consisted of think-aloud protocols with six professional content creators and an in-depth case study of iterative workflows with Design Researcher Chiara Di Lodovico.

Through these methods, the research addressed the following questions:

Main RQ:

How do AI disclosure labels shape content creators' (a) sense of authorship and (b) willingness to disclose AI use on social media?

RQ1:

How do creators interpret what AI disclosure labels imply about who made the image (authorship and credit)?

RQ2:

Which factors influence a creator's willingness to disclose AI involvement on social media?

From this thesis, it was concluded that creative agency is a relational and fluctuating process. A central conclusion of this work is that a core foundational principle for authorship resides in ideation and intent.

The findings clarify that while AI can be an assistand and a producer, the nuance (or granularity) of its involvement determines how legitimacy is perceived. Creators defend the ideation phase as the non-negotiable core of their identity. Even in scenarios where AI-generation was used beyond mere assistance, the "black box" phase was unavoidable, yet alltogether, human agency is not lost but delegated into the first and last phases. In cases where generative AI is involved, legitimacy is derived from the strategic direction and the critical evaluation provided by the human author.

Labels interact with the content they represent, therefore it is crucial to understand visualization within the broader history of visualization techniques. When photography was invented, questions about its legitimacy occured, since "pressing a button" was not perceived the same as "painting with a brushtroke" was (Bolter and Grusin, 1999). However, both mediums offered the same goal of immersing the user into an ideal of reality.

Therefore, Generative AI represents a contemporary medium that exacerbates the tension between what counts as real information and what stands for an illusion. AI further enhanced this paradigm, since the truth of an image no longer resides in its visual realism, but in its technical provenance. This matter exacerbates the collapse of traditional notions about what transparency and authorship mean.

This thesis concludes that absolute transparency can be harmful and not necessarily mean to an understanding (Ananny & Crawford, 2018), and that a the current "one-size-fits-all" approach to labels often functions as a form of hypermediacy, by interrupting users' state of flow and prompting them to negotiate and scrutinize the meaning they once only experienced. This can affect, therefore, how content and how content creators are perceived. In the findings, Theme 3 explains how some content creators use AI labels to avoid penaliza-

tions from platforms and while they accepted the practice in theory, because they value authenticity and transparency, most of them showed skepticism when they talked about their experience. For example, a participant explained fear of facing judgement for using the label, since they believed their audience would assume that they entire content was fully synthetic when they only edited a small element within the image.

Design must empower a human-centered implementation of AI by creating semiotic systems that reflect the complexity of hybrid workflows (Maverich, 2022; He et al., 2025; Burrus, 2024). Similarly, other methods to implement AI literacy could help users understand what artificial intelligence, which could lead to a deeper understanding about social media content.

Ultimately, this thesis contends that as AI becomes an contemporary productor, the role of the designer must be understood as the one of mediator of meaning.

11.1 Limitations

The stimulus set was intentionally small and context-specific, prioritizing ecological realism and depth of interpretation over exhaustive coverage of label–genre combinations. As a result, reactions cannot be attributed solely to wording, since context and aesthetics may influence interpretation. Future work should test interactive labels on content creators-

Participants were recruited from creative fields (e.g., digital art, VFX, writing/content creation) because the study investigates disclosure labels as signals of authorship, effort, and creative identity. This sampling strategy prioritizes depth in the creator perspective rather than representativeness across all social media users.

Because the sample is concentrated in creative/artistic domains, findings may not fully generalize other user groups (e.g., journalists, educators, marketers, or general audiences). Similarly, the study design prioritized ecological validity by pairing specific labels with certain contexts. While this allowed for a deep analysis of context-label alignment, it limits direct cross-comparison of the same label across conflicting domains, such as testing ‘Made with AI’ on fine art. Additionally, future research could employ a full sample to quantitatively measure using think-aloud with eye-tracking to explore how a single label’s acceptance ratings fluctuate across diverse genres.

Bibliography

12 BIBLIOGRAPHY

A

Aler Tubella, A., Theodorou, A., Dignum, F., & Dignum, V. (2019). Governance by glass-box: Implementing transparent moral bounds for AI behaviour. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 535–541). https://doi.org/10.1145/3306618.3314281

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. PNAS Nexus, 3(10), pgae403. https://doi.org/10.1093/pnasnexus/pgae403

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645

Anastasov, K. (2025, November 10). AI disclosure rules by platform: YouTube, Instagram/Facebook, and TikTok labeling guide. Influencer Marketing Hub. https://influencermarketing-hub.com/ai-disclosure-rules/

Asada, M. (2014). Towards artificial empathy. International Journal of Social Robotics, 7(1), 19–33. https://doi.org/10.1007/s12369-014-0253-z

Australian Government, Department of Industry, Science and Resources (National Artificial Intelligence Centre). Being Clear About AI-Generated Content: Spectrum of AI-Generated Content (Year of Publication/Last Update). https://www.merriam-webster.com/dictionary/document

B

Barthes, R. (1977). Rhetoric of the image. In S. Heath (Trans.), Image–Music–Text (pp. 32–51). Hill and Wang. (Original work published 1964)

Bhutani, K. (2025, March 5). SynthID: A technical deep dive into Google's AI watermarking technology. Medium. https://medium.com/@karanbhutani477/synthid-a-techni-

cal-deep-dive-into-googles-ai-watermarking-technology-0b73bd384ff6

Bickert, M. (2024, April 5). Our approach to labeling AI-generated content and manipulated media. Meta. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/

Bickert, M. (2024, April 5). Our approach to labeling AI-generated content and manipulated media. Meta. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/

Bignardi, G., Ishizu, T., & Zeki, S. (2021). Chapter 25 in Neuroaesthetics: Exploring Beauty Within and Around Us (pp. 503–504).

Boediman, E. P. (2025). Exploring the impact of deepfake technology on public trust and media manipulation: A scoping review. Jurnal Komunikasi, 19(2), 155–173. https://doi.org/10.20885/komunikasi.vol19.iss2.art8

Bomba, F., & De Angeli, A. (2025). Agency and authorship in AI art: Transformational practices for epistemic troubles. International Journal of Human-Computer Studies, 205, 103652. https://doi.org/10.1016/j.ijhcs.2025.103652

By Jason M. Allen - Colorado State Fair, Public Domain, https://commons.wikimedia.org/w/index.php?curid=122602647

C

Chalmers, D. J. (1992). Subsymbolic computation and the Chinese room. Journal of Experimental & Theoretical Artificial Intelligence, 4(3), 209–230. https://doi.org/10.1080/09528139208953747

Ciampaglia, G. L., Nematzadeh, A., Menczer, F., & Flammini, A. (2018). How algorithmic popularity bias hinders or promotes quality. Scientific Reports, 8, 15951. https://doi.org/10.1038/s41598-018-34203-2

Citterio, C. (2024, November 19). The digital author – An AI artist challenges the USCO's decision that an "AI-assisted" artwork is not eligible for copyright protection. Italy Intellectual Property Blog. https://www.ipinitalia.com/copyright/the-digital-author-an-ai-artist-challenges-the-uscos-decision-that-an-ai-assisted-artwork-is-not-eligible-for-copyright-protection/

Coleman, T. (2025, April 1). The backlash against ChatGPT's Studio Ghibli filter: The studio's charming style has become part of a nebulous social media trend. The Week. https://the-week.com/tech/chatgpt-studio-ghibli-filter-controversy

Corsi, G., Marino, B., & Wong, W. (2024). The spread of synthetic media on X. Harvard Kennedy School (HKS) Misinformation Review, 5(3). https://doi.org/10.37016/mr-2020-140

D

D. Broadbent (1958). Perception and Communication. London: Pergamon Press.

Davis, N., Hsiao, C.-P., Popova, Y., & Magerko, B. (2015). An enactive model of creativity for computational collaboration and co-creation. In Creativity in the digital age (pp. 109–133). Springer.

"Detecting Synthetic, Doubting Authentic: AI Attribution Bias for Political Imagery" Harry Yaojun Yana, Ryan C Moorea, Fangjing Tua, Jeffery T Hancocka https://share.google/rMM0X-nGkjoCM3YHxA

Detecting Synthetic, Doubting Authentic: AI Attribution Bias for Political Imagery Harry Yaojun Y

E

Earp, B. D., Mann, S. P., Liu, P., Hannikainen, I., Khan, M. A., Chu, Y., & Savulescu, J. (2024). Credit and blame for AI-generated content: Effects of personalization in four countries. Annals of the New York Academy of Sciences, 1542(1), 51–57. https://doi.org/10.1111/nyas.15258

Ericson, J. D. (2025). Reimagining the role of friction in experience design. Information Design Journal, 17(4), 131–139.

Einstein, A. (1931, February 16). Address to the student body of the California Institute of Technology [Speech]. California Institute of Technology, Pasadena, CA, United States.

European Commission. (2025, October 23). AI Act | Shaping Europe's digital future: Regulatory framework on artificial intelligence. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Parliament. (2025). Generative AI and copyright: Training, creation, regulation (PE 774.095). Policy Department for Justice, Civil Liberties and Institutional Affairs, Directorate-General for Citizens' Rights, Justice and Institutional Affairs. https://www.europarl.europa.eu/RegData/etudes/STUD/2025/774095/IUST_STU%282025%29774095_EN.pdf?utm_source=chatgpt.com

European Union (2024). Artificial Intelligence Act, Article 50. https://artificialintelligenceact.eu/article/50/

European Union. (2024). Article 5: Prohibited AI practices. In Regulation (EU) 2024/1689 on artificial intelligence (AI Act) (Chapter II). Retrieved from https://artificialintelligenceact.eu/article/5/

European Union. (2024). Article 50: Transparency obligations for providers and deployers of certain AI systems. In Regulation (EU) 2024/1689 on artificial intelligence (AI Act) (Chapter IV). Retrieved from https://artificialintelligenceact.eu/chapter/4/

F

Foucault, M. (1969). The author function (D. F. Bouchard & S. Simon, Trans.; excerpt). Michel Foucault, Info. https://foucault.info/documents/foucault.authorFunction.en/

French, L. (2025). Authorship, creativity, authenticity and originality in the media and creative industries in the age of artificial intelligence. https://doi.org/10.13140/RG.2.2.22530.39363

G

Gamage, D., Sewwandi, D., Zhang, M., & Bandara, A. K. (2025). Labeling synthetic content: User perceptions of warning label designs for AI-generated content on social media. https://dl.acm.org/doi/10.1145/3706598.3713171

Gaver, W. W., Beaver, J., & Benford, S. (2003). Ambiguity as a resource for design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 233–240). Association for Computing Machinery. https://doi.org/10.1145/642611.642653

General Data Protection Regulation (GDPR), Regulation (EU) 2016/679, 2016 O.J. (L 119) 1. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

Gibson, J. J. (1979). The ecological approach to visual perception. Houghton Mifflin. https://monoskop.org/images/c/c6/Gibson_James_J_1977_1979_The_Theory_of_Affordances.pdf

Gillespie, T. (2010). The politics of ‘platforms’. New Media & Society, 12(3), 347–364. https://doi.org/10.1177/1461444809342738

Google. (2024, December 17). Generative AI Prohibited Use Policy. Retrieved from https://policies.google.com/terms/generative-ai/use-policy

Google. (2025). Gemini Advanced Image Generation [Generative AI model]. Retrieved October 23, 2025, from https://gemini.google.com/

H

Hall https://blog.richmond.edu/watchingthewire/files/2015/08/Encoding-Decoding.pdf

Hall, S. (Ed.). (1997). Representation: Cultural representations and signifying practices (Culture, Media and Identities). SAGE Publications.

Harvard University Information Technology. (n.d.). Generative AI guidelines. Retrieved December 10, 2025, from https://www.huit.harvard.edu/ai/guidelines

Hastuti, H., Maulana, H. F., Lawelai, H., & Suherman, A. (2025). Algorithmic influence and media legitimacy: a systematic review of social media’s impact on news production. Frontiers in Communication, 10. https://doi.org/10.3389/fcomm.2025.1667471

He, J., Houde, S., & Weisz, J. D. (2025). Which contributions deserve credit? Perceptions of attribution in human–AI co-creation. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’25). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713522

I

Interaction Design Foundation - IxDF. (2020, October 8). What is Perception in UX/UI Design?. Interaction Design Foundation - IxDF. https://www.interaction-design.org/literature/topics/perception

Julien Porquet, Sitong Wang, Lydia B. Chilton (2024). Copying style, Extracting value: Illustrators' Perception of AI Style Transfer and its Impact on Creative Labor. https://arxiv.org/

pdf/2409.17410

J

Jung, Y., Hua, P., Bao, J., & Sundar, S. (2025). AI-generated or AI-modified? User reactions to labeling AI use in social media posts. Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3706599.3720264

K

Kapsetaki, M. E., & Zeki, S. (2022). Human faces and face-like stimuli are more memorable. PsyCh Journal, 11(5), 715–719. https://doi.org/10.1002/pchj.564

Khadpe, P., Wenzel, K., Loewenstein, G., & Kaufman, G. (2025). Explaining the reputational risks of AI-mediated communication: Messages labeled as AI-assisted are viewed as less diagnostic of the sender's moral character. arXiv. https://doi.org/10.48550/arXiv.2509.09645

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 2390–2395). Association for Computing Machinery.[1] https://doi.org/10.1145/2858036.2858402

Kuo, W.-H. (2025). The multimodality of generative AI internet memes in the 2024 U.S. presidential election [Doctoral dissertation, Robert Morris University].

L

Lazar, L. (2018). The cognitive neuroscience of design creativity. Journal of Experimental Neuroscience, 12, 1179069518809664. https://doi.org/10.1177/1179069518809664

Lee, H., Jung, C., Koo, N., Seo, S., Yoo, S., Hong, H., & Jang, Y. (2025). Who wants to try AI? Profiling AI adopters and AI-trusting publics in South Korea. Media and Communication, 13(2). https://doi.org/10.17645/mac.i475

Liao, R. (2024). The impact of AI-generated content dissemination on social media on public sentiment. Applied and Computational Engineering, 90(1), 82–88. https://doi.org/10.54254/2755-2721/90/20241698

Lindsey, D. (2025, April 15). A deep dive into AI labels on social media. Fuse Create. https://

fusecreate.com/a-deep-dive-into-ai-labels-on-social-media/

Lindsey, D. (2025, April 15). A deep dive into AI labels on social media. Fuse Create. https://fusecreate.com/a-deep-dive-into-ai-labels-on-social-media/

Lloyd, T., Gosciak, J., Nguyen, T., & Naaman, M. (2024). AI rules? Characterizing Reddit community policies towards AI-generated content. arXiv. https://doi.org/10.48550/arXiv.2410.11698

Lockhart, A., & Tessono, C. (2025, March). Human or AI? Evaluating labels on AI-generated social media content (The Dais Report). The Dais. https://dais.ca/reports/human-or-ai/

M

Manovich, L., & Arielli, E. (2024). Artificial aesthetics: Generative AI, Art and Visual Media. https://manovich.net/index.php/projects/artificial-aesthetics

Messer, K. D., Costanigro, M., & Kaiser, H. M. (2017). Labeling food processes: The good, the bad and the ugly. Applied Economic Perspectives and Policy, 39(3), 407–427. https://doi.org/10.1093/aepp/ppx028

Mucci, T. (n.d.). What is AI-generated content? IBM. https://www.ibm.com/topics/ai-generated-content

P

Partnership on AI. (n.d.). PAI's responsible practices for synthetic media. Retrieved from https://syntheticmedia.partnershiponai.org

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 30(3), 286–297. (Likely what was meant by "paramusam 2011").

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. Management Science, 66(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

Pilat, D., & Krastev, S. (2025). Why do our decisions depend on how options are presented

to us? The Framing Effect, explained. The Decision Lab. https://thedecisionlab.com/biases/framing-effect

Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25) (pp. 1–32). ACM. https://doi.org/10.1145/3706598.3713171

R

Radivojevic, K., Chou, M., Badillo-Urquiola, K., & Brenner, P. (2024). Human perception of LLM-generated text content in social media environments. arXiv. https://doi.org/10.48550/arXiv.2409.06653

Rae, I. (2024). The effects of perceived AI use on content perceptions. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24) (Article 978, pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642076

Ramachandran, V. S., & Hirstein, W. (1999). The science of art: A neurological theory of aesthetic experience. Journal of Consciousness Studies, 6(6–7), 15–51.

Rattner, Z. (2024). AI labels: A new voluntary content framework (Version 1.0). AI Labels Project. https://www.ailabels.org

Rob – TeamYouTube. (2024, March 18). New disclosures and labels for generative AI content on YouTube [Announcement]. YouTube Help Community. https://support.google.com/youtube/thread/264550152

S

Salma, Z., Hijón-Neira, R., & Pizarro, C. (2025). Designing Co-Creative Systems: Five Paradoxes in Human–AI Collaboration. Information, 16(10), 909. https://doi.org/10.3390/info16100909

SciPub Plus. (2025, May 20). AI-Generated Images in Academic Papers: Current Policies and Best Practices. Elsevier. https://scipubplus.com/hub/blog/ai-generated-images-in-academic-papers/

Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of undersea teleoperators. Man-Machine Systems Laboratory, MIT. (Foundational for "Levels of Automation").

Shifman, L. (2014). Memes in digital culture. MIT Press. https://doi.org/10.7551/mitpress/9780262317702.001.0001

Spillers, F. (2015, July 5). Progressive disclosure. Interaction Design Foundation. https://www.interaction-design.org/literature/book/the-glossary-of-human-computer-interaction/progressive-disclosure

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), Digital media, youth, and credibility (pp. 73–100). MIT Press.

T

TikTok. (2023, September 19). New labels for disclosing AI-generated content. https://newsroom.tiktok.com/en-us/new-labels-for-disclosing-ai-generated-content

Trattner, C., Forstner, S. L., Starke, A. D., & Knudsen, E. (2025). C2PA provenance labels increase trust in news platforms across Western countries [Manuscript in preparation]. MediaFutures, University of Bergen.

V

van Dijck, J. (2009). Users like you? Theorizing agency in user-generated content. Media, Culture & Society, 31(1), 41–58. https://doi.org/10.1177/0163443708098245

van Dijck, J. (2013). The culture of connectivity: A critical history of social media. Oxford University Press.

Verganti, R., Vendraminelli, L., & Iansiti, M. (2020). Design in the age of artificial intelligence (Working Paper No. 20-091). Harvard Business School. https://www.hbs.edu/ris/Publication%20Files/20-091_3889aa72-1853-42f8-8b17-5760c86f863e.pdf

W

Wang, N., Kim, H., Peng, J., & Wang, J. (2025). Exploring creativity in human–AI co-creation: A comparative study across design experience. Frontiers in Computer Science, 7, 1672735. doi:10.3389/fcomp.2025.1672735

Wang, Y., & Xu, D. J. (2025). Will AI replace human creators? Exploring the mechanisms of user engagement with AI-generated content on social media. AMCIS 2025 Proceedings, 8. https://aisel.aisnet.org/amcis2025/sig_svc/sig_svc/8

Ware, C. (2021). Information visualization: Perception for design (4th ed.). Morgan Kaufmann.

X

Xu, Y., Cheng, M., & Kuzminykh, A. (2024). What makes it mine? Exploring psychological ownership over human-AI co-creations. In Proceedings of Graphics Interface 2024 (GI '24) (Article 35, pp. 1–8). Association for Computing Machinery. https://doi.org/10.1145/3670947.367097

Y

Yan, H. Y., Moore, R. C., Tu, F., & Hancock, J. T. (2025, April 16). Detecting synthetic, doubting authentic: AI attribution bias for political imagery [Preprint]. OSF Preprints. doi:10.31219/osf.io/w6bzx_v1

YouTube Help. (2025). Disclosing use of altered or synthetic content. Retrieved October 12, 2025, from https://support.google.com/youtube/answer/14328491

Z

Zheng, J., Richter, A., & Hong, Y. (2025). From tool to teammate: AI as co-innovator. In PACIS 2025 Proceedings (Paper 10; PACIS2025-1260). AIS Electronic Library (AISeL).

Figures

FIGURE 1.1 & 1.2 (Social Media Ecosystem)

Van Dijck, J. (2013). The Culture of Connectivity: A Critical History of Social Media. Oxford University Press. (Adapted from p. 28).

FIGURE 1.3 (Blurry Spectrum of Content)

Australian Government, Department of Industry, Science and Resources. (2025). The spectrum of AI-generated content. National Artificial Intelligence Centre.

FIGURE 2.1 (Glass Box Model)

Aler Tubella, A., Theodorou, A., Dignum, F., & Dignum, V. (2019). Governance by glass-box: Implementing transparent moral bounds for AI behaviour. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 535–541.

FIGURE 2.2.1 (AI Act Risk Levels)

European Commission. (2026). Regulatory framework for AI. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

FIGURE 2.4 (Authorship/Attribution Diagram)

He, J., Houde, S., & Weisz, J. D. (2025). Which contributions deserve credit? Perceptions of attribution in human–AI co-creation. Proceedings of the CHI Conference on Human Factors in Computing Systems.

FIGURE 2.4 (Théâtre d'Opéra Spatial)

Allen, J. M. (2022). Théâtre d'Opéra Spatial [Digital image]. Wikimedia Commons. Public Domain. Retrieved from https://commons.wikimedia.org/w/index.php?curid=122602647

FIGURE 1.3.2 (Scheme of Perception)

Ware, C. (2003). Information Visualization: Perception for Design (2nd ed.). Morgan Kaufmann.

FIGURE 3.4 & 3.5 (Linguistic Framing & Cognitive Loop)

Medina Galán, L. (2025). Original photographs and visual synthesis.

FIGURE 4.1 (Levels of Automation)

Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of undersea teleoperators. Man-Machine Systems Laboratory, MIT.

FIGURE 4.2A & 4.2B (Label Prototypes)

Gamage, D., Sewwandi, D., Zhang, M., & Bandara, A. K. (2025). Labeling synthetic content: User perceptions of warning label designs. Proceedings of the CHI Conference on Human Factors in Computing Systems.

FIGURE 4.3 (C2PA Progressive Disclosure)

Burrus, et al. (2025). C2PA Documentation and User Study (p. 41). Cited in research.

FIGURE 4.4A (AILabels Framework)

Rattner, Z. (n.d.). AI Labels: A New Voluntary Content Framework. AIlabels.org. Creative Commons CC BY 4.0. https://ailabels.org

FIGURE 4.41B (Granular Attribution Statement)

IBM Research & He, et al. (2025). Anatomy of a granular attribution statement. Retrieved from https://aiattribution.github.io/

FIGURE 4.5A (Meta AI Info)

Instagram. (2025). UI Screenshot of AI info tag. Adapted from Meta Newsroom (Bickert, 2024).

FIGURE 4.5B & 4.7C (TikTok Labels)

TikTok. (2025). UI Prototype and existing AI-generated label.

FIGURE 4.5C & 4.7A (YouTube Interface)

TeamYouTube. (2024). Altered or synthetic content panel. Author's illustration based on YouTube UI.

FIGURE 4.5D (ArtStation #NoAI)

ArtStation. (2024). Digital asset tags for AI disclosure.

FIGURE 4.7B (Instagram Profile)

Medina Galán, L. (2025). Anonymized profile simulation based on Instagram interface.

FIGURE 8.1 & 8.2 (Stakeholders & Constraints Diagrams)

Medina Galán, L. (2025). Visual synthesis based on research findings.

FIGURE 8.4A (Spectrum of Involvement)

Medina Galán, L. (2025). Synthesis informed by: Parasuraman et al. (2000); Yannakakis et al. (2014); and Rezwana et al. (2022).

FIGURE 8.4B (Agency Curve)

Medina Galán, L. (2025). Synthesis based on interviews with Chiara Di Lodovico and Ananny & Crawford (2018).

Midjourney. (2025). Synthetic landscape and urban environment generations. Used as research stimuli.

Medina Galán, L. (2025). Natural photography food and architecture samples.

FIGURES 8.5 (Real vs. Synthetic Probes)

This section details the provenance of the visual probes used during the think-aloud sessions. The robes were categorized as "Natural Photography" (to test for false-positive AI attribution) and "Synthetic Generations" (to test for the efficacy of disclosure labels).

The following images were captured by the author using a mobile device to serve as a baseline for "real" imagery.

FIGURE 8.5a: Natural Photography with Minimalist Composition.

FIGURE 8.5c: Natural Photography (Food detail).

FIGURE 8.5e: Natural Photography with High Digital Contrast.

The following images were generated using Midjourney v6¿7.0. The prompts were designed to replicate high-fidelity photographic conventions (depth of field, motion blur, lens flare) to create perceptual realism.

FIGURE 8.5b: Synthetic Rural Landscape

Prompt: A calm, rural waterside scene on an overcast day. A gravel road starts in the

foreground and slopes gently downhill toward a lake in the mid-background. Gravel road surface mottled with small stones and slightly damp. Small houses framing the scene: red cottage with white trim on the left, yellow wooden house on the right. Still gray water and low forested hillside. High fidelity photography.

FIGURE 8.5d: Synthetic Interior Render

Prompt: Moody vintage European apartment dining nook, real iPhone photograph look, 0.5x ultra-wide angle. Large arched multi-pane window, ivy-covered courtyard wall outside. Natural overcast daylight, soft shadows, realistic window reflections. Round wooden table with wrinkled lace tablecloth, clear glass vase, mismatched wooden chairs. Worn herringbone parquet floor. Handheld snapshot imperfections: tiny tilt, mild motion blur, computational photography sharpening, natural grain.

FIGURE 8.5f: Synthetic Urban Environment

Prompt: Vertically framed street scene shot through a large, dark brick archway acting as a natural vignette. Narrow old-town lane paved with uneven cobblestones. In the foreground, a single passerby in a light black puffer jacket and tote bag, captured mid-stride with strong motion blur/defocus. Yellow ochre buildings with slatted shutters on the left, reddish-brown exposed brick with ornate metal street lamps on the right. Linear perspective, quiet and cinematic mood, sunlit hazy background.

APPENDIX A Interviews with participants (coded excerpts)

Fully anonymized transcripts are available via QR code.

SCAN ME

APPENDIX B Taguette Codebook

Artistic identity rooted in authenticity

6 highlights
uses the platform as a career tool

Openness / freedom of circulation

1 highlight
The participant prioritizes free circulation of artwork

Perceived copyright vulnerability among other artists

1 highlight

Unstable authorship boundaries in social media environments

1 highlight

Privacy and selectivity of creative process

1 highlight

AI-generated label overrides perceived human authorship

1 highlight
she sees the image and thinks it's authentic, she sees the label, and thinks it is AI generated

AI-Assisted label shows ambiguous authorship

2 highlights

AI as a threat to artistic idea ownership

1 highlight

triggered by the AI-assisted label

AI-Assisted label amplifies loss of perceived artistic authorshop

1 highlight

Frustration towards the loss of creaive agency in ideation

1 highlight
triggered by the ai-assisted

Perfection as a cue of synthetic origin

2 highlights
the visual aesthetic itself triggers the belief that the idea did not come from a human. The displacement of human authorship can occur even without the label.

conflit from unrealistic AI-generated standards

1 highlight
AI images can stablish visual expectations that exceed what is acheivable - not realistic

Disapproval of reliance on AI for creative ideation

1 highlight

Artist identity rooted in idea generation

7 highlights

AI for ideation devalues human creative capacity

1 highlight

Negotiating adoption while protecting creative agency

1 highlight

---

fear of losing the capacity to imagine without AI

2 highlights

Audience lacks critical visual discernment

5 highlights

Concern about future misuse and societal manipulation

1 highlight
is the idea is lost, identity gets disrupted, is the boundaries of what we think is real collapse then we lose truth

Ethical expectation of transparency in creative authorship

1 highlight

Labels as a practice of maintaining creator–audience trust

3 highlights

Needs labels to clarify whether a work was handmade or digitally

1 highlight
labels should prevent misinterpretation

Sees content-creator role as marginal to core artistic identity

2 highlights
Participates as a content creator but resists being defined as one,

Believes high-reach creators can exploit media to deceive

1 highlight

Audience accepts AI images as real without questioning

2 highlights
perception of audience

Audience seeks pleasurable illusion over truth

1 highlight
audience perception

Fear of a society losing critical judgment when confronted with realistic synthetic images

2 highlights

AI disrupts the ability to evaluate artistic authorship

1 highlight
when AI supplies ideas, artistic evaluation is disrupted

Authenticity relies on proving authorship of ideas

1 highlight
identity

Perceives AI-generated imagery as economically displacing illustrators

1 highlight
AI has socio-economic implications

Need for labels to show her true value over AI content

1 highlight

---

human made content is nourishing

2 highlights

Hope that human made art will be more valued for its effort and quality

1 highlight

Labels should enforce the perceived effort and quality of content

Sense of betrayal when institutions adopt AI instead of supporting human artists

1 highlight

stakeholders prioritize AI-content

Co-Creative with AI label amplifies loss or perceived human authorship

1 highlight

Curiosity towards AI as a creative tool

1 highlight

Artistic identity grounded in intrinsic personal expression

1 highlight

Art is done for self-expression and connection, not for commercial output.

Ethical awareness as creative identity

1 highlight

Curiosity towards AI in tension with his ethical commitment.

1 highlight
AI

AI for ideation is not a priority

1 highlight

AI accepted as a reference tool, but not a creative author

1 highlight
AI is allowed as long as it doesn’t threaten the identity of idea generator

AI as a scaffolding technique, not as a creative generator

2 highlights
he has strong authorship values

Lack of awareness about labels becauseAI does not an identity threat

1 highlight
AI

Sees content-creator role as a space for identity development

1 highlight

AI-generated label maintains trust and awareness in realistic but synthetic media

2 highlights
creo que está bien, porque esto se ve tan real que me vuela la cabeza de que ya hay un bien de cosas que se ven muy reales. Creo que el label está muy bueno porque te avisa con anticipación’

AI-Assisted label means a human gave idea to the machine

2 highlights

“Made with AI” label as meaning that the image has no human-originated creative input

1 highlight

Co-creation requires reciprocal effort, which AI cannot provide

1 highlight

Believes some content creators can exploit media to deceive

3 highlights

wants clarity and specificity in creative disclosure

2 highlights

process-labels or granular disclosure as necessary

Values representation of creative labor

1 highlight

Values clarity to prevent authorship and rights disputes

1 highlight

sees labeling as a practical way to reduce misunderstandings

Labels are a necessity because audiences are impressed easily

1 highlight

sees labels as segmentation cues that sort audiences by content preference

1 highlight

Sees labels as a solution to restore user agency

2 highlights

Human-made label frames creation as a competence of identities with AI

1 highlight

---

human-label marks human creativity as no longer the default

1 highlight

Human-made label risks becoming deceptive when verification is unclear

1 highlight

Self disclosure is unreliable when it competes with platform economies of attention

1 highlight

Social media dynamics do not support authenticity claims

2 highlights

---

human-made label enforces belief that the value of authenticity is already in crisis

2 highlights

AI’s impact is perceived as social rather than technical

1 highlight

---

human-made labels is a necessary approach

1 highlight

Labels are a moral practice

2 highlights

Co-created with AI label highlights ambiguity

3 highlights

Uses social media as professional showcase

1 highlight

Sees content as carrier of values (

1 highlight

Purpose-driven creativity, not trend-driven

1 highlight

3D and AI as mediums of expression, not identity threats

1 highlight

identity anchored in purpose/professionalism

1 highlight

Uses social media for professional visibility, not persona building

1 highlight

his work expresses ideas, not his inner emotional state.

1 highlight

AI explored as a technical curiosity, not as creative input

2 highlights

Professional authorship is kept separate from AI influence

1 highlight

Perceives disclosure as credibility signals

1 highlight

Sees disclosure as an industrial standard rather than a moral act.

1 highlight

Professional platforms shape his understanding of AI labeling, not social media culture

0 highlights

Authorial legitimacy depends on platform governance, not personal morality

1 highlight

Accepts AI in the workflow but rejects AI as final author

1 highlight

Labels restore perceptual clarity

1 highlight

AI-generated Label restore perceptual clarity

2 highlights

AI-generated label overrides visual judgement

1 highlight

AI-generated label transforms perception of the image's origin

1 highlight

---

fear of losing collective ability to tell real from fake

1 highlight

---

sees social media as an ecosystem where identity gets distorted. Ai is not the original problem

1 highlight

Sees social media as a space where authenticity is already unstable

1 highlight

Sees AI as a mechanism for realist simulation

1 highlight

Fear of political and harmful misuse

1 highlight

In response to social issues, labels become an urgent standard defense

1 highlight

Sees labels as fillters in an information stream

1 highlight

AI-generated image: Hyper-realism creates uncertainty

1 highlight

AI-Assisted label restores a line between image and reality

1 highlight

AI lacks intentionality and interpretation

2 highlights
AI-assisted image

AI doesn’t create, it recombines.

1 highlight

AI-Assisted marks a shift from interpretation to recombinati

1 highlight

AI triggers double standards of perception

1 highlight
anything too perfect AI,, anything too imprfect, also AI.

Creators rely on subtle visual logic to detect authenticity

1 highlight

Imperfections become diagnostic cues of fakeness

1 highlight

AI challenges creators' trust in their own visual judgment

1 highlight

Expert vs non-expert sensitivity

2 highlights

Labels help confirm what the eye cannot trust

2 highlights

Perceptual instability with synthetic images

0 highlights

---

interior design image AI output perceived as technically competent but not real

1 highlight

AI label normalizes synthetic content in professional workflows

1 highlight

The label makes AI involvement feel expected and legitimate in fields like architecture, design, interior design.

AI is already common in architectural visualization.

1 highlight

Label reduces ambiguity by matching the expected genre

1 highlight

Efficiency vs craft value tension

1 highlight

AI perfection destabilizes trust in authorship

1 highlight

AI threatens labor-based creative professions

1 highlight

AI is a threat when it replaces process, not when it replaces ideas

1 highlight

Platform labels compete with creator reputation as signals of truth.

1 highlight

Human talent is assumed real within trusted communities

2 highlights

AI-generated images collapse the boundary between talent and automation, specially in conceptual art

1 highlight

Sees co-creation with AI as incompatible with artistic intention

1 highlight

Feels co-creation exaggerates AI's artistic agency

1 highlight

Rejects co-creation because AI lacks human imperfection

1 highlight

Interprets co-creation as a literal percentage of labor.

1 highlight
"Y eso pues ya cuenta un poco como 50-50, ¿no?" He cannot understand "co-creation" unless it maps to a ratio.

Requires quantifiable creative effort to accept co-creation.

1 highlight

Questions the legitimacy of calling something co-created if the human role is minor.

1 highlight

AI obviousness makes labels unnecessary.”

1 highlight

Labels matter only when AI realism fools perception

1 highlight

AI is fine as an internal tool but not as a co-author of the artifact.

1 highlight

Fear that AI-created final outputs undermine the craft.

1 highlight

Labels protect human craft from being confused with AI output

1 highlight

Audiences will eventually sharpen their eye to distinguish human vs AI.

1 highlight

AI content is like ‘fast-food’

1 highlight

As AI becomes the norm, human-made work will require its own label

1 highlight

Revalorization of human craftsmanship

1 highlight

---

believes time and effort produce authenticity

1 highlight

---

believes people are already overwhelmed by curated, idealized realities.

1 highlight

Media performance of perfection undermines trust

1 highlight

Manipulability of narratives using AI

1 highlight

Mediatized warfare and the need for informational boundaries

1 highlight

Legal vulnerability caused by AI

1 highlight

Labels as first-line defence against misuse

1 highlight

Labels are a temporary perceptual aid, not a permanent requirement.

1 highlight

Perceived irrelevance of AI to current creative practice

1 highlight

Doesn’t consider AI as a serious tool until it becomes a clear industry norm

1 highlight

---

resists replacing his work with AI partly because of the time, money, and effort invested in learning his craft

1 highlight

Human-made label as revalorization of manual work

0 highlights

---

humans are ingenious, improvisational and abstract; AI is parameter-bound and too perfect.

1 highlight

Human authorship anchored in personal narrative rather than technical craft

1 highlight

AI accepted as an information-retrieval tool, not as a creative co-author

1 highlight

AI tools perceived as valid for factual accuracy, not emotional authenticity

1 highlight

Context-dependent emotional detachment

1 highlight

AI-assisted label amplifies perceived discomfort towards traditional craft

2 highlights

AI-assisted triggers conditonal appreciation

1 highlight

Functional acceptance of AI in practical domains

1 highlight

‘Made with AI’ label accepted under the interior design context

2 highlights

Co-created label is acceptable in fictional or world-building contexts

1 highlight

Co-created label preserves human authorship as the origin point

1 highlight

AI as a visualization tool for world-building

2 highlights

AI use accepted in low-stakes, playful, non-professional contexts

1 highlight

Co-created label seen as more legitimate and acceptable in professional environments

1 highlight

AI use in high-art formats threatens authenticity

1 highlight

Labels signal moral character and artistic integrity.

1 highlight

labels as public good

2 highlights

Mandatory enforcement increases trust in the label system

1 highlight

label secondary motivation is due to platform enforcement

1 highlight

Label use motivated by platform enforcement rather than internal moral duty

1 highlight

anticipates a future where human-made becomes a valuable

1 highlight

AI accepted for research, not creative authorship.

1 highlight

AI as an experimentation tool

1 highlight

Label reduces perceived authenticity

1 highlight

Label triggers instant disinterest

1 highlight

AI feels appropriate in functional contexts

1 highlight

“Made with AI” does not reduce value because the goal isn’t artistic authorship

1 highlight

Label Does Not Threaten Professional Identity in Non-Artistic Tasks

1 highlight

Tool-Based Recognition of Skill (“Skill in prompting still belongs to the human”)

1 highlight

Low-stakes functional images = label is normalized

1 highlight

---

authorship only matters when the image is artistic or expressive.

1 highlight

Imperceptibility of AI (“AI is invisible, indistinguishable from traditional tools”)

1 highlight

Technical Acceptance (“AI fits architectural visualization norms”)

1 highlight

Perceived AI Protagonism (“AI dominance reduces human contribution”)

1 highlight

Human Intent vs. AI Autonomy (“Human objective determines perceived legitimacy”)

1 highlight

Reluctance to Disclose Minor AI Use

1 highlight

Proportionality of AI Use (“Degree matters”)

1 highlight

AI as an edition Tool, Not a Co-Creator

1 highlight

Creator Identity Threatened by Over-labeling

1 highlight

“AI as the lazy teammate”

1 highlight

Platform automated labelling Label exaggerates AI’s contribution / “Label inflates AI credit”

1 highlight

Context-dependent acceptance of labels / “Label is fine when AI is the medium”

1 highlight

Labels protect against deception / “Labels safeguard vulnerable audiences”

1 highlight

Hyperreal AI triggers concern / “Realistic AI + label = perceptual alarm”

1 highlight

---

labels become a warning when context matters

1 highlight

Fear of losing critical judgement / ai as a threat to professional visual discernment

1 highlight

Reliance on social cues to identify AI

1 highlight

Trend-based AI recognition / “Context cues override visual cues”

1 highlight

Labels act as warning cues when the content is revelant

1 highlight

---

need for clarification in percentage, the extent AI was used

1 highlight

Creator responsibility toward the audience

1 highlight

---

proposal of percentage labeling seen as a fair practice

1 highlight

AI as a compensatory tool / “AI supports creators lacking specific skills”

1 highlight

Platform dependency for visibility (“Algorithmic pressure to produce visuals”)

1 highlight

Multiplatform creator identity

1 highlight

Artistic primary authorship (“I am an independent artist-composer”)

1 highlight

Visual content as secondary to the true creative product (“Visuals support music”)

1 highlight

AI as an assistant for workload / task delegation

1 highlight

Plural tool ecosystem / “Multiple AI tools integrated into workflow”

1 highlight

Normalization of AI in creative practice

1 highlight

High realism perception despite the label

1 highlight

‘AI-generated’ label becomes associated with possible fakery

1 highlight

---

thinks positively about the image due to detachment from visuals

1 highlight

---

utilitarian visual identity

1 highlight

AI visual generation ‘skills’ as a threat

2 highlights

‘Made with AI’ label perceived neutrally in an interior desgin context

1 highlight

---

context awareness of AI use defines the value of the content

1 highlight

‘AI-assisted’ label when AI is a creative extension

1 highlight

‘AI-Assisted’ blurs authorship boundaries enough to enable plausible deceit.

1 highlight

Creative Anxiety / “If AI can do this, what happens to human creativity?”

1 highlight

Collaborative Authorship Interpretation / “Co-created implies joint effort”

1 highlight

Co-created” feels ethically and emotionally balanced

1 highlight

Aesthetic Resistance to large labels/ “Large labels interfere with presentation”

1 highlight

---

honesty is important, but she fears the overuse of labels might distort how people judge her work.

1 highlight

AI use is okay in low stakes content

1 highlight

Context-Dependent Acceptance / “Labels are appropriate depending on the type of work”

1 highlight

Perceived Penalty for Disclosure in Artistic Fields / “Label can harm visibility and opportunity”

1 highlight

Ethical Imperative for High-Impact Content / “For news and socially impactful media, labels are mandatory”

1 highlight

“Label erases human creativity” / Fear of Misattributed Authorship

1 highlight

“Creativity still comes from me” / Assertion of human primacy

1 highlight

“Labels need nuance” / Desire for percentage disclosure

1 highlight

“Human-made label is comforting but must be aeshtetically subtle”

1 highlight

AI positive impact on production and scalability, improves work and time

1 highlight

AI increases speed, efficiency, and reduces creative bottlenecks

1 highlight

AI as Cost-Saving Tool

1 highlight

Disappointment with Human Production

1 highlight

Aesthetic documentarian rather than self-promotional creator.

1 highlight

“Creative self defined through resistant, human-sourced ideation.

1 highlight

APPENDIX C Thematic analysis

participant	code	quote	THEME	SUBTHEME
participant1	Privacy and selectivity of creative process	I rarely share my creative process or my creation process of some pieces. In general, they are always the finished pieces.	Ideation = authorship / intention	
participant1	AI-Assisted label amplifies loss of perceived artistic authorshop	I don't know if it would matter if it was digital or if it was, I don't know, an oil in a dress or a digital image. I really don't think it affects anything. I don't know. It does cause me a lot of conflict as an artist and art creator. Not just the fact that an image is produced by artificial intelligence, but that the idea is produced by artificial intelligence. I think that's the part that affects me the most as a creator, to say, let's see, we artists are not going to have,	Ideation = authorship / intention	Authorship is intention (idea, intention, interpretation, “comes from me”)
participant1	Frustration towards the loss of creaive agency in ideation	the ability to generate ideas or artistic proposals that arise from our thinking, but I have to ask the computer what happens to it, because today I don't know what to draw. Computer, well, I don't even know how to, so you can see that I'm not completely related. I hear very boomer, hey computer, I don't know how they call it, the chat GTP, I guess, right? What do I draw today? And that seems to me to be, it seems very serious, it seems very strong, very, very strong to say, I don't have the ability to say, today I'm not going to generate anything, today I'm not going to create anything because I can't think of anything.	Ideation = authorship / intention	Authorship is intention (idea, intention, interpretation, “comes from me”)
participant1	Artist identity rooted in idea generation	I don't like it. It's not the kind of work that I consume or image that I consume. Then I don't have that. Yes. I don't know, again, my conflict here is because you need the computer to give you ideas. I mean, they don't give you… Drugs are better, we use hallucinogens before a computer. I don't understand why as an artist you need a robot to tell you what to do. No, it has to be	Ideation = authorship / intention	Authorship is intention (idea, intention, interpretation, “comes from me”)
participant1	Artist identity rooted in idea generation	] I think right now I'm in a kind of resistance. To say, no, I'm not going to win. And I'm going to do it. And I'm like this. Yes, I'm in my resistance. But how long can I last?	Ideation = authorship / intention	Human-made value + authenticity as a moral stance
		I google something and the first result is the result of AI. And it gives me a lot of courage because I don't want to use AI. Why are you answering me with AI? So I always scroll through it. I prefer to go to Wikipedia or I prefer to		

SCAN ME

APPENDIX D PROBES

FIGURE 6.2.5a **AI-Generated**

Prado, A. (n.d.). Man military talking service [Photograph]. Pexels. https://www.pexels.com/sv-se/foto/man-militar-pratar-tjanst-3880204/

FIGURE 6.2.5b **Made with AI**

Google AI Studio. (2025). Modern minimalist living room interior with neutral tones and natural light [AI-generated image]. Google AI Studio.

Prompt: A professional interior photography shot of a bright, minimalist modern living room. The room features a clean off-white sofa, a large grey circular rug, and a light wood floating media console with a large flat-screen TV. A large fiddle leaf fig plant in a terracotta pot sits by a floor-to-ceiling window with soft, natural lighting.

FIGURE 6.2.5c **AI-Assisted**

Based on Night with her Train of Stars (1912) by Edward Robert Hughes. This public domain artwork was modified by the author to simulate AI-enhancement for the purpose of the study.

Hughes, E. R. (1912). Night with her Train of Stars [Painting]. Original image sourced from Unsplash. Digital modifications/AI-enhancements by the author (2025).

FIGURE 6.2.5d **Co-Created with AI**

Google AI Studio. (2025). Mystical glowing portal in a bioluminescent forest with floating islands [AI-generated image]. Google AI Studio.

Prompt: A digital fantasy painting of a glowing neon-blue circular portal centered in a dark, ancient forest. The portal reveals a surreal world featuring floating islands, a massive celestial tree, and distant planets. The foreground is filled with bioluminescent mushrooms, gnarled roots, and ethereal floating jellyfish creatures.