Google

Determining trustworthiness through context and provenance

December 2024

Summary

As tools powered by generative AI become more accessible and widespread, debates about the ability for people to determine the trustworthiness of media content have become more prevalent and urgent. These concerns are not new, but rapidly-progressing generative AI capabilities can present new challenges.

In this paper, we consider these concerns about synthetic content within the broader context of information literacy. Research tells us that users are primarily concerned with determining whether they can trust the information they encounter in their online journeys. In light of recent developments in generative AI, many suggest that information about the way content was created (by AI or otherwise) can be a proxy for assessing trust. However, "Is this AI-generated?" is not equivalent to "Is this trustworthy?". Though these two questions can overlap, additional context is often needed to make decisions about trustworthiness.

Accordingly, we will examine the value and limitations of both **assertive provenance** and **inferred context**, both of which play an important role in empowering users to make sound decisions on content they encounter:

- **Assertive provenance** refers to techniques used at the creation or editing stage to provide a clear signal regarding the means of creation of a piece of content (artificially generated, edited, etc.) – for instance, watermarking, fingerprinting, or metadata.
- **Inferred context** refers to techniques that aim at sussing out important information about a piece of content without relying exclusively upon information included by its creator or editor – for instance, leveraging information available on the open web to infer where a piece of content came from, how old the content may be, what claims are being made about it and by whom, or how the content is being used in other contexts.

Ultimately, both techniques can play an important role in helping people make informed decisions about what to trust online. We believe that a holistic solution is needed, comprising assertive provenance tools, in-product context tools, and off-platform investment in information literacy capacity building programs. To be successful, this requires a multistakeholder approach, uniting industry, civil society, governments, academic experts, and users in a collaborative effort to develop and refine the tools and programs necessary for maintaining the integrity of our information ecosystem.

Foreword

It has never been more critical to be able to evaluate the trustworthiness of information we see online.

Nowadays, the volume of information we consume and the speed at which it is presented to us can feel daunting at best; it can feel even more challenging as technology rapidly evolves. Discerning trustworthy information is necessary not only to understand what we see and hear online, but to meaningfully participate in and contribute to an ever-changing society.

Educators and librarians have been doing this work for a long time – preparing students to find, evaluate, and understand information effectively through new media and technological advancements. While fact checking organizations, journalists and information literacy experts have developed best practices to discern the trustworthiness of information, many people feel unequipped to practice these tactics when they need them. In fact, seven in ten (70%) respondents to a 2023 study led by the Poynter Institute's digital media initiative MediaWise reported not being totally or very confident in their ability to tell when online images are authentic and reliable [1].

A common question we hear is how to tell if an image was AI-generated. While this is certainly an important question, understanding whether something is real or AI-generated does not always help us understand the trustworthiness of the larger narrative or claim being made. That is, asking "Is this AI-generated?" is not the same as asking, "Can I trust this?" For example, an image may not be AI-generated but may still be taken out of context or manipulated with photo-editing software. There are other, broader credibility questions that people should ask, such as: Is this image being used in the right context? Where did this information come from? What is the perspective or incentive of the person who is sharing it? Is there a bigger picture to consider? And even more challenging, in many cases answering the question "is this true?" is complicated and does not have a clear answer.

People come to Google to verify information they see elsewhere - maybe it's a text message from a family member or something shared on social media. Our approach to helping people find more information is two-fold: first, we fundamentally build our products from the ground up with quality in mind. That means when people come to Google they will readily find high quality information they can trust. Second, we believe people should have access to easy-to-use tools that provide the context they need to help them answer the question "can I trust this?" for themselves– even when they aren't on Google. We do this by building tools that leverage the best of Google to help people understand the credibility and context of something they're seeing online. Our tools and features don't require any advanced technical skills to use, and they are built to plug into how people live their day-to-day lives.

Finally, we don't build these products just based on what we alone think will work. There will always be new ways for people to create and consume content. We know that we must keep learning and listening to the information literacy community and the people who use our products in order to continuously evolve and improve our solutions. We engage deeply with the information literacy community to understand how people interact with the changing information ecosystem. We learn from the research of scholars (some of which is included in this paper), as well as from new best practices and interventions in the field. This helps us understand the best ways to approach this ever-evolving challenge within our products and ensures that our tools help people strengthen their own information literacy skills. This essential partnership enables us to take a leading role in providing best-in-class solutions in collaboration with the larger ecosystem.

Our work in this space is far from done. At the speed of technological change, something that works today may no longer work next year, and we understand that technological solutions alone are not sufficient either. We're committed to working with the larger community to continuously evolve and update our approach in this space, and are eager to learn from the discussions that may be sparked by the findings highlighted in this paper.

1 Introduction

From digital safety to digital literacy

As tools powered by generative AI become more accessible and widespread, debates about the ability of people to determine the trustworthiness of media content have become more prevalent and urgent. These concerns are not new, but rapidly-progressing generative AI capabilities can present new challenges. This calls for updated approaches to assess content trustworthiness, including a combination of assertive provenance and inferred context, in addition to broader information literacy efforts.

When new technologies emerge, the discourse about them naturally includes concerns over risk and abuse – and methods evolve to best protect everyday consumers. Accordingly, well-intentioned "digital safety" efforts aim to arm users with the skills to understand the risky content they may encounter. For example, advice to combat phishing and email scams in the early days of the web initially centered on practical tips like closely examining unfamiliar senders or steering clear of dubious links. However, these methods were neither foolproof nor sustainable: scammers adapted, and they quickly learned to use spell check and secure .org domains (which had previously been touted as a marker of trustworthiness) [2, 3].

We face an analogous scenario with AI-generated content. The general public is awash with well-intended advice to separate the "real" from the "fake", often recommending that users look out for anomalies in digital images. Today, everyday users are told to check for extra fingers and arms, blurry or abnormal backgrounds, overly smooth surfaces, accessory mistakes, garbled text, or inconsistent shadows [4–7]. This list-based approach focuses on what may be practical for busy users, at the expense of what is consistently reliable and – more importantly – durable. These tips have largely been developed to aid users with visual markers that may work in some cases today, but may not work at all in the future as technology improves [8, 9].

The false allure of easy heuristics

This type of advice, often packaged as "AI literacy" or "digital literacy," focuses on providing users with easy heuristics, ones which seemingly anyone can follow. However, these are not consistently reliable and can even lead to conspiratorial-minded thinking about whether a shadow pattern in a given photo really is possible or whether the use of photo editing software indicates some deeper truth that is being hidden from us [10, 11].

The desire for quick shortcuts for determining trustworthiness is understandable, especially given that on average, people are not adept at determining whether an image has been manipulated. Moreover, we are even worse at identifying what has been altered in a photo, if it has indeed been changed [12]. Given all of the above, the need for evolved information literacy strategies is clear, especially as the public encounters a higher volume of synthetic content.

This paper, therefore, considers the emergence of generative AI capabilities within the broader, existing framework of information literacy and trustworthiness. It focuses on:

1. **Drawing the foundational distinction between questions like "Is this AI-generated?" and "Is this trustworthy?"**
2. **Understanding the landscape of proposed information literacy tools, both via assertive provenance and inferred context**
3. **Outlining the benefits and downsides of assertive provenance solutions**
4. **Outlining the benefits and downsides of inferred context tools**
5. **Understanding user-facing labels**
6. **Recommending a holistic approach to information literacy, situated within the existing evidence base on misinformation and disinformation**

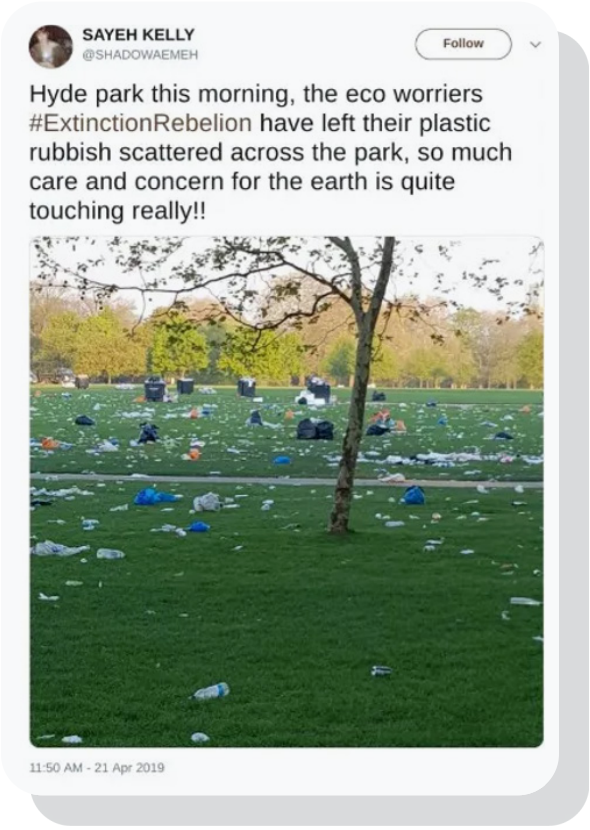
2 Determining trustworthiness

A rich body of evidence already exists around the strategies and tactics that are effective in addressing whether content is trustworthy or not [8, 13, 14]. But before even considering the best approaches to helping the general public make informed decisions about what to trust, we first must step back and determine whether we are asking the right set of questions in the first place.

Is this trustworthy?

Generative AI has undeniably revolutionized content creation, offering unprecedented levels of realism, creativity and efficiency [15], but it has also sparked concerns about the potential for deception [16, 17]. Here it is important to ask the right questions when considering trustworthiness.

For both traditional and synthetic images, the critical question focuses on whether the content itself – and the claims surrounding it – can be trusted. For every viral deliberately misleading synthetic image [18], there are thousands of deliberately misleading photos that rely upon more simple tactics, ranging from adding a misleading caption, to cropping, re-coloring, or re-touching [19–21]. The first photographic "hoax" involved no file or forensic manipulation at all, only clever marketing [22]. And one of the most common forms of visual misinformation remains the same: "real" photographs taken out of context, such as photographs from floods in Bangladesh [23], and a still from a movie erroneously described as a police van stolen in a riot in France [24]. In a study of publicly fact checked images, context manipulations were the single largest type of manipulation, totaling 55% [25].



The critical question, therefore, is not just whether a given piece of content is generated by AI, but whether it is trustworthy. Sometimes these two questions – "Is this AI-generated?" and "Is this trustworthy?"– overlap or complement one another; sometimes they do not. Accurately assessing content's trustworthiness often requires more than a check of how it was originally created – it demands a deeper understanding of context surrounding an image [16, 26, 27].

"[This image] truly shows rubbish in Hyde Park. However, the image was taken in the aftermath of a celebration for 420, a marijuana-centric holiday, not a global-warming protest." Source: *snopes.com*

Prompt “hyde park in summer with trash everywhere”
Source: Google AI

Prompt “art of hyde park london with trash.”
Source: Google Imagen

Consider the examples above. The first is an image depicting litter in Hyde Park, London. It claims to show the aftermath left by climate activists and is held up as evidence of their hypocrisy. In fact, the photograph is real but miscontextualised: it was taken after a 420 celebration, a day associated with marijuana use [28].

The second is an AI-generated image created from the prompt “hyde park in summer with trash everywhere”, and depending on the context, it would be more or less trustworthy than the real photograph. If it was published in a tweet like the first example, with the text “Hyde park this morning, the eco warriors #ExtinctionRebellion have left their trash everywhere”, it would be considered untrustworthy. However, if it was used during a litter picking campaign to illustrate what the park could look like if it was covered in trash, it could be considered more trustworthy.

The third is an AI-generated image created from the prompt “art of hyde park london with trash.” It is clearly artistic, rather than photorealistic. The fact it was generated by AI doesn’t make it less trustworthy.

While the content of these images is similar, their trustworthiness is not – and this determination hinges not exclusively on their “realness,” but on the context within which they are situated and the claims they purport to make.

Inferred context vs. assertive provenance

The example above shows the importance of assessing content in its broader context, beyond the binary of AI-generated or not [29, 30]. Two sets of approaches have surfaced to facilitate this: assertive provenance and inferred context. These approaches enhance our digital literacy toolkit, enabling an in-depth understanding of the context, origins, and integrity of a piece of content.

Assertive provenance employs a range of techniques to mark the nature and source of a piece of content as it is created or edited, for example, whether AI has been used to generate or significantly alter it or by whom. Such methods can include watermarks, fingerprinting, metadata, or other indications that can signal the authenticity of a piece of content. As beneficial as assertive provenance can be, it also presents its own set of challenges, as we will explore.

Inferred context empowers users to deduce the credibility and authenticity of content through context clues, historical data, or the application of digital forensic tools. This set of methods encourages active engagement and critical thinking, allowing for a more nuanced assessment of content credibility utilizing multiple sources. Yet, this approach also presents challenges, primarily the need for a certain level of skill and knowledge to effectively employ the techniques, potentially placing an undue burden on people as part of their online journeys.

Both assertive provenance and inferred context offer a unique set of advantages and challenges when applied in discerning the reliability of content online. Below, we consider these concepts, examining the pros and cons of each, and how they can be utilized.

3 Understanding assertive provenance

Assertive provenance involves the explicit marking of content to signal its origins and edits. The methods that can be used to that end predate the rise of generative AI and are used in many other contexts; they include:

- **Watermarking:** Embedding digital markers into the content. The watermark can be either visible or invisible to the human eye. Non-visible information stored in a watermark can be retrieved by a detector specific to that watermark.
- **Fingerprinting:** Converting a piece of content and relevant information about it into a compressed sequence of numbers (the ‘fingerprint’) that is stored in a database that can be searched later, in order to verify whether a new piece of content matches.

- **Markup and metadata:** Including information about a piece of content in its metadata, such as its date of capture, generation or editing. This technique is particularly valuable when it comes to cryptographically signed metadata - that is, metadata that is difficult to edit without readers being aware a change has been made.
- **Declaration:** Collecting information from content creators that provides transparency on how content may have been created and/or edited.

Advantages of assertive provenance

The use of watermarking, fingerprinting, markup, and metadata to signal content provenance can help users distinguish between content that has been captured by a camera or a smartphone and content that may be partly or wholly synthetic. The advantages of these techniques include:

- **Provenance techniques help provide enhanced transparency and accountability:** By embedding digital watermarks or fingerprints or providing relevant metadata, content creators, platforms or digital developers can offer a transparent account of how content was generated and the tools that were employed [31]. This kind of transparency provides a signal that content distributors or consumers can rely upon, because it is deliberately included with the content at the time of creation and travels with it online. Accordingly, companies such as Google have started to deploy such techniques and empower creators on their services to do the same [32].
- **Provenance information can support consumer decision-making:** The application of clear, understandable user-facing labels conveying provenance information can help equip people with the knowledge necessary to make informed choices about content they consume [33], especially in cases where the risk of harm is high. This is particularly relevant when distinguishing AI-generated from human-created content becomes increasingly complex.
- **Growing collaboration on provenance across the content creation and distribution ecosystem are rife with potential:** Some provenance techniques such as cryptographically signed metadata are very interoperable and increase in utility as they are adopted across a range of content creators, producers, or distributors. One such example is the Coalition for Content Provenance and Authenticity (C2PA), whose provenance metadata standard ("content credentials") has a fast-growing community of contributors and implementers including but not limited to technology companies like Google, OpenAI, and Microsoft; media companies like the BBC and CBC; camera companies like Sony; and more.
- **Platforms can integrate provenance information into their products and policies:** Provenance information can be integrated into product design, similar to how Google Search and YouTube are including IPTC and/or C2PA metadata in user-visible features. Google Ads systems are also starting to integrate C2PA signals to help enforce key policies.

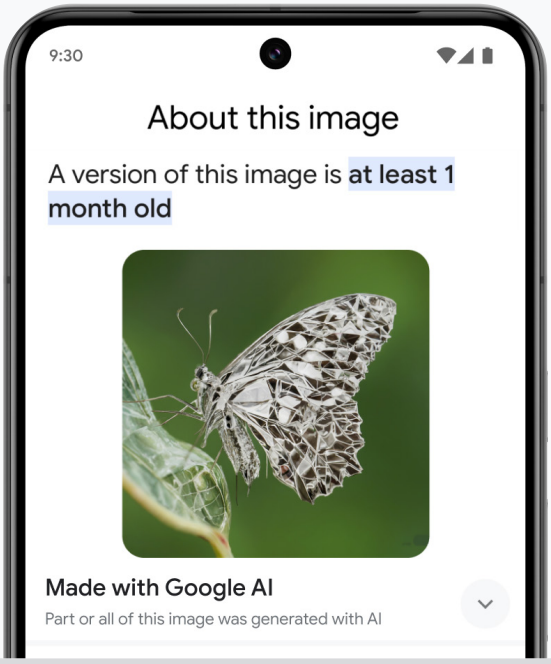
Assertive provenance at Google

Google is integrating assertive provenance along with a suite of other features to allow users to check whether a piece of content is generated by one of its AI consumer products. The development of these tools form part of Google's wider efforts to advance AI responsibly in alignment with Google's AI Principles, and to engage with various forums across government, industry, and civil society.

Assertive provenance features at Google include SynthID, which watermarks and identifies AI-generated content by embedding digital watermarks directly into AI-generated images, audio, text or video. SynthID watermarks are imperceptible to the human ear and eye, but detectable for identification. The watermark is robust to a wide range of image manipulations, and remains detectable even after modifications like cropping, adding filters, changing colors, changing frame rates (for video) and saving with various lossy compression schemes (commonly used for JPEG images). Forgery or removal of the watermark is difficult but technically possible.

SynthID is an extension of Google's existing integration of assertive provenance features into its products. Google Images, for example, already had IPTC metadata integrated into its results. Now, IPTC metadata is also available for imagery generated by Google products. In addition, the "About this image" tool detects SynthID watermarks and displays that information within Image Search. Together, these features provide people with more information about all images generated by Google AI consumer products.

Google believes that investment in information literacy is a multifaceted endeavor that must encompass the wider technology ecosystem in order to be effective. To this extent, we open sourced SynthID for text to developers to enable them to use this technology to help them detect whether text outputs have come from their LLMs. Google is also a steering committee member of the Coalition for Content Provenance and Authenticity (C2PA) and has recently joined the International Press Telecommunications Council (IPTC) as a Voting Member.



Limitations of assertive provenance

At the same time, assertive provenance is not a panacea. From watermarking to labeling to metadata, all have certain drawbacks. These include both technical and non-technical limitations.

There is no existing technical solution that can detect all synthetic content and then watermark or label it appropriately. The technical limitations of assertive provenance approaches fall into categories of feasibility, scalability and abuse risks. These can include the following:

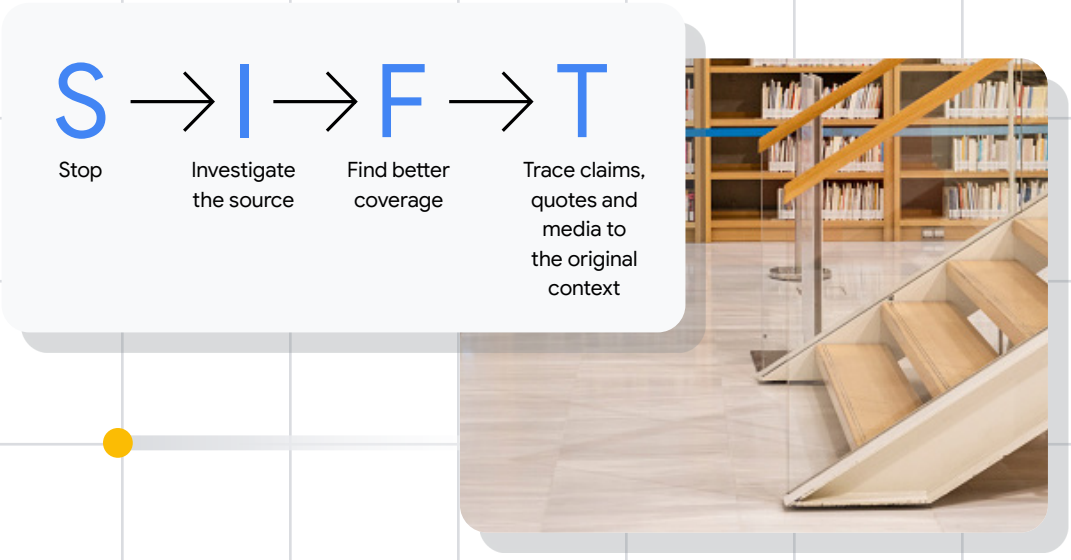
- **Watermarks, even when tamper-resistant, can be removed:** Imperceptible watermarks can be made tamper-resistant and, thus, remain with an image even if cropped or resized. However, a determined person can bypass watermarking through a number of techniques, which require varying degrees of technical skill [34]. Further technical research is required to address some of these existing challenges [35]. When a watermarked image has been merged or added to another image, it may be confusing for a detection tool to determine which part of the image the watermark relates to.
- **Metadata may be edited or stripped:** Metadata can be helpful in providing information about the origin and history of digital content. However, IPTC metadata can be edited or removed using common photo editing software. C2PA metadata, meanwhile, stores information through cryptographic signatures that verify the integrity and authenticity of the content, so it can be removed but not modified. Nonetheless, even tamper-resistant metadata can be stripped from media during format conversions [36].
- **Reliance on responsible actors, combined with a lack of incentives for bad actors:** While watermarks and labeling solutions provide critical tools for verifying content, their effectiveness fundamentally relies on their application by responsible actors. Bad actors have no incentive to use tools that include assertive provenance methods like watermarking, mark-up, or metadata – and they can easily use derivatives of open-source software or models to create digital content without any of these safeguards.

4 Understanding inferred context

Inferred context refers to gathering contextual information to trace the history, context and origin of a piece of content. This relies on digital skills, competencies and critical thinking to assess credibility. Key questions to consider include:

- **When did the image or claim first appear?** Identifying the timeline can help determine if the content is being reused or repurposed in a misleading context.
- **Where did it come from?** The source(s) of the content can significantly influence its reliability. Knowing more about the background and credibility of a source can help people make more informed decisions about its trustworthiness.
- **Why was it published?** Knowing the perspective and background of the creator or distributor of the content can provide insights into motivations or potential biases.
- **What do other sources say about it?** Cross-referencing information with other reputable sources can validate content accuracy or expose discrepancies.

Several methods have been developed to allow users with different skills to assess the credibility of online content and information. These include the COR (Civic Online Reasoning) method developed by Sam Weinberg at Stanford University, and the SIFT method, conceived by digital literacy expert Mike Caulfield. The SIFT method is a concise strategy that advocates for users to "stop" before engaging with content, followed by steps to "investigate the source", "find better coverage", and "trace claims, quotes, and media to their original context". This emphasizes critical thinking and verification through lateral reading and corroborating evidence.



Advantages of inferred context

By studying the successful practices and habits of professional fact checkers, information literacy techniques like SIFT focus on training everyday users to trace and corroborate claims in a more neutral, process-oriented way. Lateral reading, for example, is aimed at finding other coverage on a given topic or source; it helps avoid the confirmation bias or reliance on unhelpful heuristics that can lead users to inaccurate conclusions.

Other advantages of leveraging inferred context in evaluating trustworthiness include the following:

- **Empowering critical engagement:** Inferred context encourages users to actively engage with content through a critical lens, fostering a deeper level of scrutiny. By asking key questions about first appearance, source, creator, and corroboration with other sources, users are not just passive consumers, but active evaluators of information and their own assumptions. This is essential in an era where users frequently encounter online dis- and misinformation [37, 38].
- **Enhancing digital and information literacy:** By using the SIFT framework users enhance their ability to navigate the complex digital information landscape with discernment. The questions help develop users' digital literacy competencies [37] and promote the development of essential digital literacy skills.
- **Promoting healthy skepticism and verification practices:** Inferred context instills a healthy skepticism towards online content, urging users to verify information before accepting it as truth. This is crucial in an information ecosystem where authenticity can be masked or manipulated. The methodology behind inferred context, including reverse image searches and the evaluation of metadata, equips users with the tools to assess content credibility from multiple angles.
- **Providing scalability and applicability to a wide range of content:** Inferred context approaches can be applied to any piece of content in any context, provided there is enough publicly available information about it. Unlike assertive provenance, inferred context does not require good-faith participation of every AI developer or content creator within the ecosystem. It can also be applied to content that has been manipulated by using more simple tactics, such as adding misleading captions or photoshopping. This scalability across a wide range of content can mitigate some of the risks that labeling from assertive provenance approaches are susceptible to (e.g. the implied truth effect, discussed below in section 5).

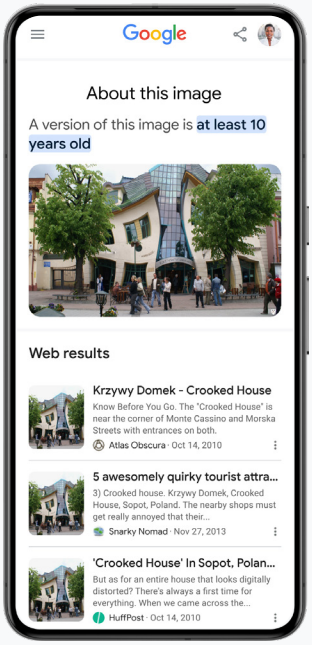
Building inferred context tools at scale

Building on evidence-based practices, user research and input from information literacy experts, several Google tools are available to users globally to facilitate inferred context techniques. These tools enhance users' abilities to infer the context and reliability of digital content and can provide critical insights for informed decision making:

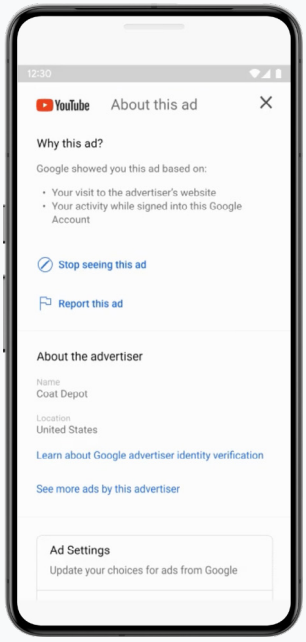
"About this result" and "About this page" provide background and perspectives on the source of a webpage and multiple perspectives on the topics covered, encouraging lateral reading and verification across multiple sources.

"About this image" allows users to assess the context and reliability of an image by highlighting the history of the image, how other sites use and describe the image, and the image's metadata, when available.

"About this ad" allows users to see information on why particular ads are being shown to them. In the disclosure, users may also see the name and location of the advertiser behind a selected ad as well as a link to the advertiser's page in the Ads Transparency Center, where they can see more ads by the advertiser.



Citation: "The crooked house in Sopot" by Topory, licensed under CC BY-SA 3.0



On YouTube information panels may appear on certain videos relating to topics prone to misleading information, such as the moon landing. These panels provide additional information and context on the topic, sourced from independent third party partners.

In addition, if a channel is owned by a news publisher that is funded by a government, or publicly funded, an information panel providing publisher context may be displayed on the watch page of the videos on its channel. And when a user watches a YouTube video on a health-related topic, they may see an information panel providing context on the source underneath the video. This panel is designed to help users better understand the sources of health content on YouTube, and to encourage lateral reading.

Limitations of inferred context

Inferred context offers a robust framework for assessing the credibility of digital content, but it also has specific limitations and challenges that can impact its effectiveness. These include:

- **Dependence on user skills and competencies, as well as time and willingness to dive deeper:** Effective inferred context heavily relies on users' digital literacy, critical thinking, and familiarity with digital tools. Understandably, not every user has the expertise or time to effectively navigate and interpret the wealth of information available online. This disparity can lead to inconsistent application and effectiveness of inferred context techniques across different user groups, leaving some more susceptible to misinformation.
- **Reliance on the broader information ecosystem:** While numerous tools and resources support inferred context, their efficacy can vary if content that debunks a false claim simply does not exist yet. For example, discerning the credibility of an image may be limited if a given image is new and hasn't yet appeared in other pages across the open web. Similarly, lateral reading on a given topic may not be effective if a false or controversial claim is new and reputable organizations like news or fact checking sites have yet to publish articles on the topic.

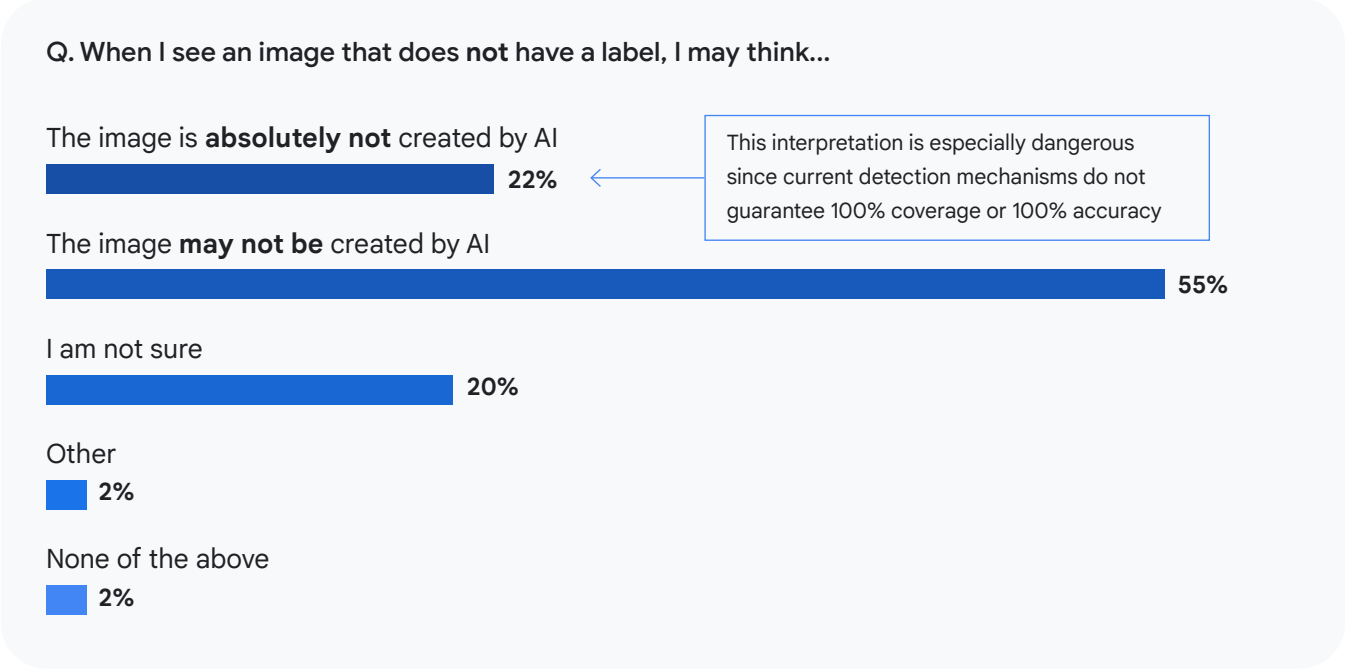
While understanding the context of a piece of content is valuable in helping people make informed choices about what is trustworthy, its effectiveness relies on users having the skills, motivation, and time to engage in information literacy practices. In an era when misinformation is sticky and attention spans are low, it is important to make inferred context practices as easy and as frictionless as possible.

5 Understanding user-facing labels

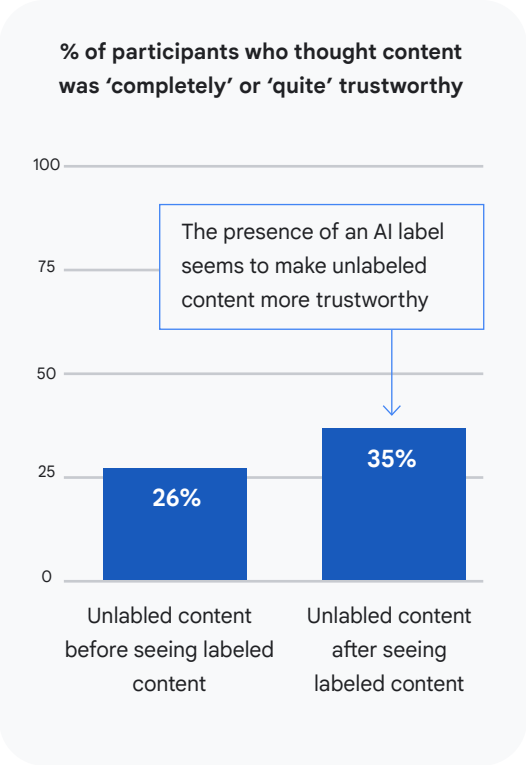
In an information environment clouded with mis- and disinformation, the application of clear, understandable user-facing labels conveying contextual or provenance information can help equip people with the knowledge necessary to make informed choices about the content the consume. This is particularly relevant when distinguishing AI-generated from human-created content becomes increasingly complex.

However, **user research has demonstrated additional risks of prominent, user-facing labels**. While labels can provide meaningful context to a user, a labeling strategy that is not applied responsibly can unintentionally lead to the following:

- **Users falsely believing any unlabeled content is "real":** While labels can provide valuable cues about content provenance, their effectiveness is contingent upon users' ability to accurately interpret and trust these indicators. In a study [39] with approximately 2700 U.S. participants, 22% believed that unlabelled content was "real", when in actuality the lack of a label means it is unknown how an image was made, not that the image is "real". This result accords with what some misinformation researchers have called the "implied truth effect" – a phenomenon by which labeling certain content as "false" may lead others to assume unlabelled content has been verified and is true [36].



- **Users falsely believing that unlabeled content is more trustworthy:** Similarly, labeling some content as generated or altered by AI and not labeling other content can lead users to the mistaken belief that unlabeled content is more trustworthy. In fact, this is not the case; rather what the unlabeled state indicates is that it is unknown whether the content was generated or altered by AI, not whether it is "real", nor whether it is trusted. This approach can inadvertently lead to damaging and harmful false conclusions. In another study [40] with approximately 5600 U.S. participants, 35% of those shown unlabeled content that followed labeled content indicated high levels of trust, compared to only 26% high trust responses for those shown unlabeled content alone. These results suggest the labeling caused a 9% increase in high levels of trust—an increase that is not justified by the information presented to respondents. In this case, AI labels seemingly caused an illusory inflation in trustworthiness of unlabeled content.



- **Reduced trust in content creators that responsibly apply content labels:** Because AI detection systems are not yet comprehensive or 100% accurate [41, 42], labeling proposals often rely on self-disclosure by the creator of the content. However, providing a user-visible label indicating that content has been altered can mean that responsible creators who label innocuous images correctly as AI generated can suffer a loss of trust. Again, the lack of a label does not mean content is "real," just that its origin is unknown. Labeling some content but not others could therefore lead to a situation where those who do responsibly self-disclose end up less trusted than those who do not.
- **User-visible labels applied liberally could lead to banner blindness:** Finally, as the ecosystem evolves and more and more content is made by generative AI tools, a ubiquitous labeling strategy could lead to banner blindness. User responses to labels are influenced by selective attention, meaning they may not always heed or even notice labels over time [43–45]. Banner blindness has been studied in the context of cookie consent fatigue [46], and lessons learned in this settings should also inform labeling strategies in the context of synthetic content. Simply put, because of limited attention spans, users tend to focus on their specific information needs, rather than on every pop-up, ad, or banner that crosses their screen.

In summary, while labeling can provide users with context, its effectiveness is contingent upon a range of factors, including user understanding, psychological impacts, and the broader context of information literacy. Addressing the limitations associated with labeling must balance technical solutions with educational initiatives and critical engagement to foster a more informed and resilient digital society.

6 Conclusion

Discerning the trustworthiness of digital content is and has always been complex, and it defies one-size-fits-all solutions. It is a multifaceted issue that necessitates a holistic approach, blending assertive and inferred context with wider educational efforts that focus on AI and information literacy.

Since simply asking "Is this generated by AI?" does not suffice in assessing content trustworthiness, it is necessary to develop a broad and varied toolkit. Here it is important to draw upon the extensive history of digital literacy, including efforts to tackle the issues surrounding online information and content [47]. This means employing digital tools, not just for identification, but for thorough evaluation and analysis of digital information [48–51]. Digital literacy demands we consider the context within which digital information exists.

Focusing predominantly on technical measures like watermarking and metadata may lead to an overreliance on these solutions, potentially neglecting the critical role of user education and information literacy. Assertive provenance, while valuable, is not a panacea; it must be complemented by efforts to enhance users' ability to critically evaluate content, understand the nuances of digital information, and navigate the information ecosystem with discernment. Conversely, neglecting technical measures to solely anchor on user education and information literacy would miss out on the opportunity to leverage important signals that can help users make more informed decisions and distributors build better services.

Toward a Holistic Solution

Both assertive provenance and inferred context serve as vital components of a broader digital literacy toolkit, each playing distinct but complementary roles. Building on our experience and third party research, we believe that a holistic solution would comprise four key areas:

Best Practices	Examples
Assertive provenance tools where there is confidence in their reliability and comprehension.	SynthID, Creator disclosures
In-product context tools to make information literacy easier for everyday consumers.	About This Image, Prebunking, Information panels on YouTube, About This Ad, Creator disclosures
Partnerships and cross-industry efforts to help provide more transparency and context for users.	C2PA, IPTC
Off-platform investments in evidence-based information literacy programs.	Super Searchers, Be Internet Awesome, Prebunking

As it becomes more challenging for people to make informed decisions on what to trust due to increased changes in digital content, the integration of assertive provenance and inferred context, together with information literacy capacity building efforts, will be crucial.

To be successful, this requires a multistakeholder approach, uniting industry, civil society, governments, academic experts, and users in a collaborative effort to develop and refine the tools and programs necessary for maintaining the integrity of our information ecosystem.

While it is not possible to devise a "silver bullet" for discerning trustworthiness, it is possible to forge a comprehensive, adaptive toolkit that empowers users to engage with digital content critically and responsibly. By accepting the complexity of this challenge and committing to a collaborative, multifaceted, multistakeholder response, we can aspire to a digital environment that enriches and informs, supporting our collective desire toward a more informed and discerning global society.

Acknowledgments

Thank you to Paree Zarolia, Joe Paxton, Jenn Shreve for conducting the user research studies that were crucial to the development of this paper. Thank you to Nidhi Hebbar, Clement Wolf, Isabelle Stanton, Paul Haahr, Ryan Kelly, Alyona Miliutina, Sherif Hanna, Katharina Familia Almonte, Chris Bregler, Ernesto de la Rocha Gomez, Ross Bower, Geoff Samek, Jess Hemerly, Sven Gowal, Armin Senoner, Sam Greenfield, Beth Goldberg, Miriam Estrin, Nick Bauer, Camino Rojo, Raquel Vazquez, Aparajitha Vadlamannati, Meghann Farnsworth, Christa Muldoon, Eve Novakovic, Roxanne Carter, Sarah Dudley, Katherine Harrison, Rebecca Shapiro Duggan and the whole team dedicated to building and improving the tools at Google that support both assertive provenance and inferred context. Thank you to Harris Cohen and Daniel Rocha for your early support. And thank you to Zoe Darme, Sebastian Smart, Charles Bradley and Mevan Babakar for synthesizing all these insights and developing this paper.

Finally, thank you to the many researchers who are cited in this paper – and who are expanding our understanding of evidence-based practices to address misinformation, disinformation, information integrity, and information literacy.

References

[1] Poynter Institute for Media Studies. A Global Study on Image Information Literacy. August 2023, https://www.poynter.org/wp-content/uploads/2023/10/A-Global-Study-On-Image-Information-Literacy-August-2023.pdf (August 2023, accessed 1 November 2024).

[2] Button M, Nicholls CM, Kerr J, et al. Online frauds: Learning from victims why they fall for these scams. Australian & New Zealand Journal of Criminology 2014; 47: 391–408.

[3] Breakstone J, McGrew S, Smith M, et al. Why we need a new approach to teaching digital literacy. Phi Delta Kappan 2018; 99: 27–32.

[4] Hern A. How to tell if an image is AI-generated. The Guardian, 8 April 2024, https://www.theguardian.com/technology/2024/apr/08/how-to-tell-if-an-image-is-ai-generated (8 April 2024, accessed 1 May 2024).

[5] Steele C. Can You Spot AI-Generated Images? Take Our Quiz to Test Your Skills. PCMAG, https://www.pcmag.com/articles/how-to-detect-ai-created-images (2024, accessed 1 May 2024).

[6] Rowe. 9 Simple Ways to Detect AI Images (With Examples) in 2024. tech.co, https://tech.co/news/ways-detect-ai-images-examples (2023, accessed 18 April 2024).

[7] Potter B. Spotting AI-Generated Images: Tell-Tale Signs and Skill-Testing Photo Examples. Medium, https://medium.com/@brittanynpotter/spotting-ai-generated-images-tell-tale-signs-and-skill-testing-photo-examples-f31c9655f418 (2023, accessed 1 May 2024).

[8] Dame Adjin-Tettey T. Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education. Cogent Arts & Humanities 2022; 9: 2037229.

[9] Qian S, Shen C, Zhang J. Fighting cheapfakes: using a digital media literacy intervention to motivate reverse search of out-of-context visual misinformation. Journal of Computer-Mediated Communication 2022; 28: zmac024.

[10] Cramer J. Settling the Controversy Over Photo of Lee Harvey Oswald. Darmouth University, https://home.dartmouth.edu/news/2015/10/settling-controversy-over-photo-lee-harvey-oswald (2015, accessed 18 April 2024).

[11] Rogers. The Kate Middleton Photo's Most Glaring Photoshop Mistakes | WIRED. Wired, https://www.wired.com/story/kate-middleton-photoshop-mistakes/ (2024, accessed 18 April 2024).

[12] Caldwell S, Gedeon T, Jones R, et al. Imperfect Understandings: A Grounded Theory And Eye Gaze Investigation Of Human Perceptions Of Manipulated And Unmanipulated Digital Images. Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science (EECSS 2015). Barcelona, Spain – July 13 - 14, 2015, Paper No. 308

[13] Domenico GD, Sit J, Ishizaka A, et al. Fake news, social media and marketing: A systematic review. Journal of Business Research 2021; 124: 329–341.

[14] Fard AE, Lingeswaran S. Misinformation Battle Revisited: Counter Strategies from Clinics to Artificial Intelligence. In: Companion Proceedings of the Web Conference 2020. Taipei Taiwan: ACM, pp. 510–519.

[15] Valyaeva A. AI Image Statistics: How Much Content Was Created by AI, https://journal.everypixel.com/ai-image-statistics (2023, accessed 5 March 2024).

[16] Das M, Fiannaca A, Morris MR, et al. From Provenance to Aberrations: Image Creator and Screen Reader User Perspectives on Alt Text for AI-Generated Images. CHI Conference on Human Factors in Computing Systems, https://research.google/pubs/from-provenance-to-aberrations-image-creator-and-screen-reader-user-perspectives-on-alt-text-for-ai-generated-images/ (2024).

[17] Cao Y, Li S, Liu Y, et al. A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT, http://arxiv.org/abs/2303.04226 (2023, accessed 5 March 2024).

[18] Marcelo P. FACT FOCUS: Fake image of Pentagon explosion briefly sends jitters through stock market. AP News, https://apnews.com/article/pentagon-explosion-misinformation-stock-market-ai-96f534c790872fde67012ee81b5ed6a4 (2023, accessed 1 May 2024).

[19] Fanfani M, Iuliani M, Bellavia F, et al. A vision-based fully automated approach to robust image cropping detection. Signal Processing: Image Communication 2020; 80: 115629.

[20] Bik E. Opinion | Science Has a Nasty Photoshopping Problem. The New York Times, 29 October 2022, https://www.nytimes.com/interactive/2022/10/29/opinion/science-fraud-image-manipulation-photoshop.html (29 October 2022, accessed 18 April 2024).

[21] Karaszewski L, Alexander S. Why a manipulated magazine photo plays a pivotal role in 'The People v. OJ Simpson'. LAist, https://laist.com/shows/the-frame/why-a-manipulated-magazine-photo-plays-a-pivotal-role-in-the-people-v-oj-simpson (2016, accessed 18 April 2024).

[22] Zhang M. The First Hoax Photograph Ever Shot. PetaPixel, https://petapixel.com/2012/11/15/the-first-hoax-photograph-ever-shot/ (2012, accessed 18 April 2024).

[23] Times Fact check. FAKE ALERT: Photos of boy rescuing fawn from drowning are not from Assam floods. The Times of India, 24 July 2020, https://timesofindia.indiatimes.com/times-fact-check/news/fake-alert-photos-of-boy-rescuing-fawn-from-drowning-are-not-from-assam-floods/articleshow/77145349.cms (24 July 2020, accessed 1 May 2024).

[24] AFP Australia. This photo shows a street parade in Switzerland before the COVID-19 pandemic. Fact Check, https://factcheck.afp.com/photo-shows-street-parade-switzerland-covid-19-pandemic (2020, accessed 1 May 2024).
[24] Novak M. Viral Photo Of Stolen Police Van In France Actually From Netflix Movie. Forbes, https://www.forbes.com/sites/mattnovak/2023/07/02/viral-photo-of-stolen-police-van-in-france-actually-from-netflix-movie/ (2023, accessed 1 May 2024).
[25] Dufour N, Pathak A, Samangouei P, et al. AMMeBa: A Large-Scale Survey and Dataset of Media-Based Misinformation In-The-Wild. Epub ahead of print 21 May 2024. DOI: 10.48550/arXiv.2405.11697.
[26] Bharati A, Moreira D, Flynn PJ, et al. Transformation-Aware Embeddings for Image Provenance. IEEE TransInformForensic Secur 2021; 16: 2493–2507.
[27] Moreira D, Theisen W, Scheirer W, et al. Image Provenance Analysis. In: Sencar HT, Verdoliva L, Memon N (eds) Multimedia Forensics. Singapore: Springer Singapore, pp. 389–432.
[28] Evon D. Were Piles of Rubbish Left in Hyde Park By Global-Warming Protesters? Snopes, https://www.snopes.com//fact-check/protesters-hyde-park-rubbish/ (2019, accessed 18 April 2024).
[29] Fazio L. Out-of-context photos are a powerful low-tech form of misinformation. The Conversation, http://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959 (2020, accessed 11 March 2024).
[30] Brennen S, Simon F, Nielsen R. Beyond (Mis)Representation: Visuals in COVID-19 Misinformation. The International Journal of Press/Politics 2021; 36: 277–299.
[31] Ecker UKH, Lewandowsky S, Tang DTW. Explicit warnings reduce but do not eliminate the continued influence of misinformation. Mem Cogn 2010; 38: 1087–1100.
[32] Gowal S, Kohli P. Identifying AI-generated images with SynthID. Google DeepMind, https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/ (2023, accessed 11 March 2024).
[33] Morrow G, Swire-Thompson B, Polny JM, et al. The emerging science of content labeling: Contextualizing social media content moderation. Asso for Info Science & Tech 2022; 73: 1365–1386.
[34] Zhao X, Zhang K, Su Z, et al. Invisible Image Watermarks Are Provably Removable Using Generative AI, http://arxiv.org/abs/2306.01953 (2023, accessed 22 May 2024).
[35] Srinivasan S. Detecting AI fingerprints: A guide to watermarking and beyond. Brookings, https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/ (2024, accessed 1 May 2024).
[36] Cvet M. Examining C2PA Provenance Metadata in DALL·E 3 Images. Medium, https://mikecvet.medium.com/examining-c2pa-provenance-metadata-in-dall-e-3-images-64ed51159091 (2024, accessed 1 May 2024).
[37] Chakroff A, Cole R. Provenance information reduces intent to share subtly manipulated images. Preprint, PsyArXiv. Epub ahead of print 26 July 2022. DOI: 10.31234/osf.io/jptv4.
[38] Bowles J, Croke K, Larreguy H, et al. Sustaining Exposure to Fact-checks: Misinformation Discernment, Media Consumption, and its Political Implications. SSRN Journal. Epub ahead of print 2023. DOI: 10.2139/ssrn.4582703.
[39] Google UXR Study> 27,000 participants
[40] Google UXR Study> 5,600 participants
[41] Thompson SA, Hsu T. How Easy Is It to Fool A.I.-Detection Tools? The New York Times, 28 June 2023, https://www.nytimes.com/interactive/2023/06/28/technology/ai-detection-midjourney-stable-diffusion-dalle.html (28 June 2023, accessed 1 May 2024).
[42] Kovtun D. Testing AI or Not: How Well Does an AI Image Detector Do Its Job? bellingcat, https://www.bellingcat.com/resources/2023/09/11/testing-ai-or-not-how-well-does-an-ai-image-detector-do-its-job/ (2023, accessed 1 May 2024).
[43] Stewart DW, Martin IM. Intended and Unintended Consequences of Warning Messages: A Review and Synthesis of Empirical Research. Journal of Public Policy & Marketing 1994; 13: 1–19.
[44] Argo JJ, Main KJ. Meta-Analyses of the Effectiveness of Warning Labels. Journal of Public Policy & Marketing 2004; 23: 193–208.
[45] Hancock PA, Kaplan AD, MacArthur KR, et al. How effective are warnings? A meta-analysis. Safety Science 2020; 130: 104876.
[46] Wein T. Data Protection, Cookie Consent, and Prices. Economies 2022; 10: 307.
[47] Gilster P. Digital literacy. New York: Wiley Computer Pub., http://catdir.loc.gov/catdir/toc/onix04/96046961.html (1997, accessed 5 March 2024).
[48] Martin A. A european framework for digital literacy. NJDL 2006; 1: 151–161.
[49] Ng W. Can we teach digital natives digital literacy? Computers & Education 2012; 59: 1065–1078.
[50] Kim KT. The Structural Relationship among Digital Literacy, Learning Strategies, and Core Competencies among South Korean College Students. EDUC SCI-THEOR PRACT 2019; 19: 3–21.
[51] Churchill N. Development of students' digital literacy skills through digital storytelling with mobile devices. Educational Media International 2020; 57: 271–284.