
HOLDING THE LINE: Authentication, Verification, and the Fight for Facts in the AI Age

Princeton CITP

Authors:

Hilke Schellmann, Sohyeon Hwang, Marianne Aubin Le Quéré,
Felix Chen, Sophie Luskin, Arvind Narayanan, Mihir Kshirsagar;
with expertise from Aimee Rinehart and Swapneel Mehta.

*Special thank you to all of the workshop participants for their
meaningful conversations from which this report was borne.*



SUMMARY

Almost anything can now be generated with artificial intelligence: photographs, audio, video, documents, entire websites. As synthetic content becomes cheaper and more convincing, and as verifying what we see online grows harder, how do we sustain an information ecosystem in which facts can still be established and trusted? In this report, we identify the challenges brought to the information ecosystem by the increased accessibility of generative AI. We start from the premise that verification, authentication, and transparency are key pillars of online information integrity. We then describe how this emerging ecosystem challenges the trustworthiness of facts from the perspective of information producers, consumers, and intermediaries.

Next, we turn to a discussion of potential interventions that can address the challenge of AI-mediated information integrity. These interventions are not exhaustive, but they represent the issues raised by participants drawn from a cross-section of journalists, researchers, and technologists who participated in an in-person workshop held at NYU's Arthur L. Carter Journalism Institute in early June 2026.

This report is intended to serve as a bridge across different actors in this complex information ecosystem: newsroom engineers, investigative and open-source reporters and editors, verification tool builders, independent journalists, policy makers, researchers, and media platforms. We aim to surface the primary challenges — and imagined interventions — to work towards a more resilient information future together.

1 Introduction: the challenge of trustworthiness

As synthetic content becomes cheaper and more convincing, and as verifying what we see online grows harder, how do we sustain an information ecosystem in which facts can still be established and trusted? Two futures are visible from here. In one, we lose confidence in facts altogether and settle into a world of rumor and motivated speculation. In another, a smaller set of trusted producers hold their ground while the surrounding flood of unverified material is met with appropriate suspicion. Neither is guaranteed, and both may coexist. This report traces potential paths to address the integrity crisis and is a call to action for greater collaboration and conversation across the information ecosystem.

The economics of information work against careful reporting on three fronts. Faking content is cheap, and getting cheaper. Verifying is time-consuming and will likely get more time-consuming soon as the reference points a reporter once checked against — a website, a document, a public record, a voice on the phone — can now themselves be fabricated, so establishing one fact increasingly means establishing the facts beneath it. While AI tools can strengthen the verification process and introduce new opportunities to increase access to knowledge, there is an arms race with the volume of potentially fabricated facts. Finally, dismissing something true has become nearly free: anyone can call a genuine report fake or machine-made, what is termed the “[Liar’s Dividend](#),” and the accusation does its work without having to be true.

Left without easily accessible and dependable ways to tell facts from falsehood — audiences drift toward whatever is plausible or suits their inclinations and is available, and over time that pressure risks pushing careful work out of circulation. This is occurring against a larger societal retreat from forming relationships with news organizations and [trusting them](#), and the resulting strain on democratic discourse. A core challenge ahead is how do we build trust in facts when the institutions that produce them are also underresourced, under siege, and less recognizable to audiences.

The proliferation of generative AI content across the internet in recent years has brought the deeper crisis of declining trust in the information ecosystem into stark [focus](#). Starting from the early 2000s, with the rise of social media and the internet, information production and sharing became broadly available to the public and enabled new actors such as bloggers to bypass institutional gatekeepers to post content online. Whereas traditional markers of trustworthiness (i.e., that information would be reliable, have integrity, and producers could be held to account if wrong) had been tied to institutional producers of information like newsrooms or TV networks, the task now falls to audiences to assess the credibility of all sorts of information providers. These conditions set the stage for an online ecosystem unprepared for the epistemic implications of generative AI, where content production is easy, cheap, and scalable, while the costs of information accuracy and verification are expensive and time-consuming.

This report draws together what a day-long workshop of journalists, technologists, policymakers, and academics surfaced about the problem of information integrity in the AI Age. It proceeds in three sections: the mechanisms societies use to establish trust; the incentives of producers, consumers, and intermediaries in the information ecosystem; and a set of potential interventions.

Our goal is to push forward the conversation of how to make tomorrow's information ecosystems more resilient and trustworthy.

2 Authentication, Verification, and Transparency

It is worth disentangling three mechanisms that are sometimes placed under the umbrella of fact-checking to describe the process for establishing that a given piece of information is trustworthy: authentication, verification, and transparency. We discuss each in turn.

2.1 Authentication

Authentication, or content provenance, is the mechanism of confirming that a piece of information is real. This concept parallels the legal evidentiary standard from U.S. [Federal Rule of Evidence 901](#), which requires "evidence sufficient to support a finding that the item is what the proponent claims it is", for example, establishing that a photo is indeed an authentic photo and not generated. Importantly, authenticating a document does not establish that the information it contains is true. Take an [image of a missile remnant](#) outside a school in Israel – how do we know that the photograph is real? This requires extensive work on the part of the investigator to authenticate the image. But even if the image is real, that does not tell us if the missile remnant shown is, in fact, a fragment of an Iranian missile.

The challenge for journalists and online investigators today is that images can be easily faked. Especially when relying on images circulating on social media that are not taken under a known chain of custody, investigators need to develop new mechanisms to confirm the source. In lieu of this institutional trust, some newsrooms have turned to technical approaches such as cryptographic keys and digital signatures that are adapted from the field of computer and information security. One of the leading open standards for establishing the provenance of digital content is the Coalition for Content Provenance and Authenticity (C2PA), which is supported by the cloud software company Adobe, large technology companies including OpenAI, Amazon and Meta, and various newsrooms.

Implementing C2PA. C2PA attaches an encrypted metadata manifest to all content shot with C2PA-enabled cameras. The manifest embedded in the file's metadata will either update with edits or generate an additional manifest based on the significance of the changes. It aims to establish a chain of custody. The technology has been described as "glass-to-glass," describing the ideal goal of an unbroken, cryptographically secure chain of custody — from the glass in the camera to the glass on a consumer's cellphone or computer.

One news organization, which is pioneering the use of C2PA, shared some of the implementation details. On the technical side, the news organization relies on third-party software to ingest, edit, and publish photos (and other media). The ingestion software at first stripped the C2PA manifest from the digital files, and the news organization had to engage the third-party vendor and convince them to address the problem. They then worked with their archiving software vendor, which had recently changed its product model to a cloud-based system. With the switch to the new system, they first thought that the C2PA manifests were stripped from the files. It turned out it was not, but implementation of a standard like this, which touches every step of the workflow, requires ongoing testing and retesting. It also showed the news organization how vulnerable they are to third-party product changes.

As of writing this report, the news organization has been able to keep the C2PA manifest intact for in-house use, but the content management system (CMS) vendor is not supporting C2PA as of now, although they are in communication with the vendor and are hoping for a change. It seems that most CMS vendors will need to update their systems if they want to support C2PA.

So for now, the manifest gets stripped from photos and other digital files before publishing — which could pose a trust problem. Will audiences believe the news organization has a chain of custody for its files if it cannot offer "full transparency" of the manifest(s)? Would audiences essentially just have to take the news organization's "word" for it? Or would the newsroom weaken its own long established trust signal if it introduces another credential? Would an icon or checkmark work in place of the full manifest(s), and how should news organizations educate readers and general audiences about the process and what such checkmarks or

icons might mean? It seems that for national legacy news organizations with a trusted name this new standard may weaken their own brand, but for many local news organizations that run national or international news, this might bolster trust if audiences understand the standard.

In the implementation process, privacy and surveillance concerns also came up. Publishing the full manifest(s) of a C2PA-enabled photo may itself constitute a privacy violation in some jurisdictions, or may cause harm by inadvertently leaking location information about an anonymous source or the reporter. Privacy and surveillance concerns become even more complex in hostile environments: authoritarian regimes could track reporters via the C2PA manifest, which is able to capture additional metadata about the camera.

Many news organizations also rely on freelance reporters or social media sources who may not have C2PA built into their systems. How might audiences react to images that are unverified but posted in the news organization's products? The same issue arises when news organizations send photographers into hostile environments and, for fear of surveillance, deliberately equip them with non-C2PA-enabled devices — which would then appear in the news organization's products as unverified images alongside authenticated, C2PA-enabled ones. We recommend that trailblazing news organizations share their C2PA implementation experiences, so that other organizations thinking of doing the same, can learn early on and don't have to face the same or similar implementation hurdles.

Many of the signatories and companies affiliated with the C2PA standard are large technology companies, whose interests in supporting the standard may not always align with those of newsrooms. For verification purposes, newsrooms often rely on videos and images posted to social media platforms, which regularly strip metadata from content. Verification experts have long called for this metadata to remain attached to the content, but the social media companies represented on the steering committee have chosen not to implement that solution. As with any standard that depends on third-parties, there is also a risk that companies could scale back their support or change their products over time, potentially leaving newsrooms without the backing they relied on.

2.2 Verification

Verification is the second mechanism of establishing that information is trustworthy. This is the process of triangulating information to confirm the [facts](#). Authenticating an image through a mechanism like C2PA can only confirm that the image was shot by a certain camera and how it was edited, but it doesn't show if what was captured is actually truthful. Returning to [the image of an alleged missile remnant](#), the question is whether the image, even if real, is actually that of an Iranian missile that fell in the school on the day in question. The New York Times story describes how the journalists verified the image by talking to experts to confirm the origin of the missile fragment.

The verification challenges already in play are a preview of more structurally disruptive threats. When trying to verify if the information in an image is trustworthy by calling experts and eye witnesses, how can a reporter be confident that the person they are interviewing over Zoom or by phone is the person they say they are? How can they verify that a website genuinely belongs to the organization it claims to represent, when AI can generate convincing institutional facsimiles — functional websites, plausible domain histories, fabricated organizational records — at minimal cost? And how will this more challenging fact-checking environment play out as generative AI could soon degrade the underlying information supply?

Traditional verification and fact-checking has long relied on cross-referencing websites, documents, and digital traces. Each of those reference points is now subject to synthetic manipulation. This raises a harder foundational question: what counts as an authoritative source in journalism? The category has always been somewhat underspecified, but it was stabilized in practice by the relative difficulty of fabrication. Books are rarely fact-checked. Websites are becoming unreliable. Human experts could be avatars, bots, or fake. As the cost of generating convincing institutional content falls, the profession needs a more explicit and institutionally legible account of what it actually means to establish a fact.

On the other side, AI and machine learning tools can help researchers and journalists verify information and dig through copious amounts of data to find interesting stories. For example, many news organizations used such tools when reviewing materials from the Epstein files to break stories.

In the broader digital ecosystem, crowdsourcing dynamics demonstrate how verification happens outside of traditional institutional settings. For example, on English Wikipedia, volunteer editors work together to report and update

articles on breaking news topics in real time, such as [natural disaster crises](#) or ongoing [protests](#). In recent years, platforms have integrated crowdsourced verification as features such as [Community Notes](#) after cutting ties with [fact-checkers](#). [Research](#) suggests that Community Notes could reduce the virality of posts and [increase](#) the likelihood of incorrect content being deleted by the original poster. At the same time, Community Notes can also [contribute](#) to patterns of information disorder online by enabling bad actors and trolls to provide and “verify” misleading information. Signals of verification can thus become subverted.

2.3 Transparency

Transparency is the third mechanism that is crucial to establishing trust. This mechanism aims to make the process and interests driving information production visible to the audience. A number of newsrooms now use “How We Reported This” sidebars, link directly to raw source documents (via Document-Cloud), or use the reporting process for example in open-source intelligence (OSINT) investigations as a narrative thread of the journalistic piece. It would be helpful to have newsrooms with special visual/OSINT investigations teams publish how they verify content step by step and what counts as reliable source, so other, less resourced newsrooms, may be able to learn from their approach.

We would also urge newsrooms to build transparency mechanisms — around reporting, standards, fact-checking, source selection, and topic selection, which is now mostly happening in the background of reporting — into the work they produce, rather than assuming news consumers already understand what goes on behind closed doors. Sharing reporting methods, particularly in investigative work, can also help; the podcast *In the Dark*, especially the third season, offers one example worth pointing to.

3 Drivers of Information Integrity

In this section we analyze what drives different actors in the information ecosystem to invest in the adoption of the three mechanisms we discuss: authentication, verification, and transparency. Looming in the background are deeper questions of trust in institutions that have traditionally served as stewards of factual and civic information, and are now coming under siege. Kharazian and Hill (2026) lay out how public knowledge institutions have come under coordinated attacks that undermine their legitimacy or perceived reliability. Although questions of trust and legitimacy are outside the scope

of this report, we recognize the importance of trust in how mechanisms of authentication, verification, and transparency respond to and interact with these broader societal challenges that are shaping the fight for facts in this era.

3.1 Information producers

Every producer — from editor, reporter, producer in legacy news organizations to news influencers — faces interrelated constraints of money, labor, and time, and these constraints bear directly on whether authentication and verification happen, how carefully, and for how long. Authentication and verification requires additional steps, specialized expertise, technical tools and databases, and sometimes dedicated staff, which mostly only the largest and well-funded media organizations can afford — all of which compete for finite resources.

Producers therefore concentrate on information they believe will be most impactful, where “impact” is defined by their goals: civic value, community benefit, virality, or ad revenue. The broader political economy of the media shapes these choices as most producers depend on some mix of advertising, subscriptions, and platform reach to survive, and different funding models generate different pressures. In addition, the competition for attention in online environments tends to reward shareable, catchy content. And that pressure is felt unevenly across different producer types.

Newsrooms. Legacy newsrooms and the reporters within them are the producers most explicitly organized around fact-checking and verification. Professional values do much of the work here: journalists are motivated by goals to follow the facts, keep the public informed, and “speak truth to power,” and these values narrow attention toward civically important events. These values are reinforced through training and editorial oversight, and they are operationalized through explicit infrastructure: style guides, attribution standards that govern sourcing like the “two-source rule,” and correction practices that create accountability when fact-checking fails. Crucially, in newsrooms these standards serve overlapping functions — ensuring internal consistency, protecting accuracy, and limiting legal liability. But the values can also fail, especially under resource pressure, competitive speed, deference to powerful sources, or thinly-staffed fact-checking (often entry-level or freelance, and easily overridden) can all erode the verification the values are meant to guarantee.

It’s important to note that most newsrooms’ outputs, including articles, pod-

casts, videos and even social media posts, are subject to standards, (extensive) editing, and in some instances legal review and fact-checking by a designated fact-checker (for most breaking news and reporting on deadline, the reporters themselves are asked to fact-check their work). But most of this takes place internally before publication and many readers and audience members have no idea that such a process exists to ensure that the content is factually correct. We have seen that some legacy newsrooms are breaking the fourth wall and are opening up their decision making and policy processes, but it might be helpful to include that in the actual articles akin to the visual investigations at *The New York Times* (see: [Behind the Journalism: How The Times Works](#) and [How The Times Assessed That Photo From Trump of Maduro in Handcuffs](#)).

News Influencers. The growing cadre of news influencers turn the reporting and fact-checking process inside out and make them [part of their work](#), actually building trust with audiences. These independent creators are diverse actors on different social media platforms shaping different outputs. They are dependent on audience engagement and often put that front and center. They also have to build trust from the ground up since they cannot rely on institutional credibility like “legacy” news organizations. “What we find is that individual creators often work hard to demonstrate ability, benevolence, and integrity to build trust with their audiences. They narrate their expertise, respond to reader questions or suggestions, and interact with their critics — all tactics that help build trust,” explains Julia Angwin in her [2024 analysis](#) of news influencers. Legacy news rooms have much to learn from such actors.

3.2 Audience

On the audience side, an entirely different set of features drive whether information is trusted and believed to be accurate. In this section, we synthesize a few observations about how audiences validate and decide whether information is trustworthy, particularly in online contexts.

Online-native credibility heuristics. The amount of information we encounter on a daily basis exceeds our ability to verify each claim – people therefore tend to rely on rapid heuristics or signals to make decisions about information quality online. For example, when gauging whether information online is reliable, people often rely on quick signals like search ranking, social engagement signals, and professional-looking website design. While relying on these signals is an understandable response to a dense information environment, online heuristics do not always accurately reflect whether or not information

has been verified or authenticated. For example, generative AI can enable anyone to easily create professional-looking websites and text, and bots make it easier to fabricate social signals. It is thus key that any interventions address not only how authentication, verification, and transparency happen, but also how they can be communicated easily to audiences.

Institutional trust. Institutional trust is one of the key ways people attribute trustworthiness to content. In short, when people trust an institution, they also attribute trust to the content that institution produces. For example, people may believe content created by an institution like Reuters or the CDC has inherent value and quality. Recent studies from Pew suggest that institutional trust in traditional media is broadly in decline, and generally much more so for conservative audiences. People are turning to non-institutional actors to make decisions and quality judgment. The issue is not that institutional trust no longer works; rather the beneficiaries of this trust have shifted.

Tech-enabled provenance tools. As technology companies invest in the technical development of authentication, verification, and transparency solutions, parallel research seeks to understand how audiences might respond to these signals. There does not, as yet, appear to be a clean mapping between AI disclosure transparency and audience trust: on the contrary, disclosing AI use tends to decrease audience trust in news articles. However, the parsing of provenance signals appears to be highly contextual, for example AI disclosure for financial reporting or on sports matches is perceived less negatively than on political topics. Ultimately, more research is needed to understand how to maintain trust with audiences while engaging in appropriate AI disclosure (which is also a moving target among news organizations and news influencers).

3.3 Intermediaries

Intermediaries such as social media networks sit between producers and the audience and control the distribution of information. But the incentives of these platforms rarely align with authentication or verification. At their core, they are driven by engagement and advertising revenue. They have little inherent stake in the reputational value of authenticated, truthful information because their business models monetize attention, regardless of whether the underlying content is a verified fact or a synthetic fabrication.

Platforms do occasionally intervene to support reliability — such as banning certain types of deepfakes, introducing labels, or downranking known disinformation. However, these actions are typically defensive, driven by brand

safety concerns, advertiser pressure, or the threat of regulation, rather than a structural commitment to information integrity. Reliability is often treated as a compliance metric or a PR necessity, not a core product value. There are periodic legislative efforts to reform Section 230 of the 1996 Communications Decency Act to remove the litigation shield that has protected platforms from taking the distribution of reliable information more seriously, but those efforts have not succeeded to date.

The historical relationship between intermediaries and information producers complicates this dynamic further. Over the past two decades, platforms have operated on an extractive model, effectively hollowing out the business models of legacy newsrooms by capturing the majority of the market of digital advertising that relies on the journalists' content to keep users engaged. Because platforms have historically captured the value of reporting without bearing the cost of its production, the challenge of the AI age goes beyond building better authentication tools. The core question is how to shift the underlying incentives of distribution: how do we compel intermediaries to actively support the distribution of high-quality, verified content?

Finally, we note the rise of a new information intermediary — AI chatbots — that have a different set of incentives at play that we intend to discuss in greater detail in a separate paper. For present purposes, it's worth noting that spitting out an incorrect or unsupported answer can be harmful to the audience and so the models have, in theory, an incentive to get the answer right. But the AI models face two challenges — they risk severing the attribution loop, killing the click-through that funded the producers whose work they draw on, and the answer interface risks collapsing the work of careful provenance and verification into an unsourced assertion provided as a neutral overview.

4 Potential interventions

Supporting clearinghouses for best practices. As listed in our accompanying newsroom guide, many news organizations now employ a variety of technical tools to help with the authentication and verification process. It takes a lot of time for individual reporters to surface the newest tools, how to use them and which tools have changed or disappeared entirely. Supporting collaborative clearinghouse mechanisms, such as those discussed below, for continuously testing these tools and sharing reporting tips could help all reporters - in large and in small newsrooms.

Involving journalists early in tool development. Many reporters noted that many of the current authentication tools' confidence measures do not translate to real-world reporting. A 70% confidence rating that an image is not AI-generated is not helpful for reporters, since they are more comfortable with a binary assessment about whether an image is authentic or not. These confidence ratings are not (easily) translatable to audiences, since they want to know if an image is genuine or not and not a confidence rating that reporters often cannot even explain since the authentication tools are often "black boxes" themselves and are not transparent why an image, for example, got a 70% confidence rating. We urge technologists to work with reporters to develop the tools together and then let reporters evaluate their usefulness.

Developing humanness checks. During the workshop participants observed that there is often no reliable way for journalists and others to verify whether a human is on the other end of a Zoom or phone call. There are some technical measures, but those are not always feasible for a journalist approaching a source. Journalists need reliable ways to confirm that the source they are speaking to is in fact human and we urge technologists to work on a solution here.

Building collaborative verification infrastructure. Most current interventions operate at one of two scales: individual journalist effort (verification checklists, detection tools) or platform-level content moderation. Shared verification infrastructure — analogous to, say, the role wire services once played in pooling reporting capacity — remains underdeveloped. But there are promising models like the former First Draft (now [Information Futures Lab](#)), the [Global Investigative Journalism Network](#), and [Bellingcat's](#) open-source intelligence methods that could be scaled up.

Using AI tools wisely. [AI tools](#) can help journalists cross-check and verify factual claims, especially when managing large datasets. Similarly, AI tools used by scientists to reproduce research can also help reporters test scientific claims and explain complex concepts in simpler terms. These tools have important limitations, but they can advance the process of verification.

Prioritizing consumer education. Almost all current interventions target better tools for producing and certifying trustworthy content. However, technical tools can also help make complex verification information more legible: what would a literacy intervention look like — how can complex verification concepts be explained more clearly? One that takes audience skepticism seriously as partly warranted while still creating pathways back to institutional credibility? Too often, the tools are developed by technologists without consulting end users whether they are the reporters or the audiences.

Encouraging transparency mechanisms. Current tools, to the extent they work today, are better at verifying content — whether an image is authentic and/or has been manipulated — than at verifying source, whether the person presenting it is who they claim to be. Closing that gap will require developing privacy-preserving mechanisms for transparency that don't reproduce the risks C2PA has exposed, alongside clearer transparency mandates for AI-generated content itself.

Leveraging digital intermediaries without creating a lasting dependence on them. Despite their extractive relationship, digital platforms offer a unique mechanism for a newsroom to reach their readers in more meaningful ways beyond clicks on a website. Digital creators are a global example of how to create and maintain a direct and lasting connection with audiences. There are lessons newsrooms can learn from intermediaries, to regain some of the value they have lost to the latter, without creating complete dependence on them.

AUTHORS



Hilke Schellmann



Sohyeon Hwang



Marianne Aubin Le Quéré



Felix Chen



Sophie Luskin



Arvind Narayanan

WITH ADDITIONAL EXPERTISE FROM

DESIGN & COPYEDITING



Mihir Kshirsagar



Aimee Rinehart



Swapneel Mehta



Lydia Owens