WITNESS

C2PA Content Credentials and the Surveillance Risk:

Adversarial Scenarios and Governance Gaps in the Content Provenance Ecosystem

Verifying the provenance of digital content, including whether and how AI was involved, is essential to rebuilding trust in what we see and read online. But the infrastructure being built to do that must not become a tool for surveillance and control.

ACKNOWLEDGMENTS

With thanks to WITNESS colleagues: Sam Gregory, Mahsa Alimardani, Bruna Martins dos Santos, shirin anlen, Zuzanna Wojciak.

This report also benefited from the time, expertise, and critical engagement of the people listed below, who were invited to review and scrutinize the scenarios and analysis. Their input sharpened the work considerably. Their inclusion here does not imply agreement with the findings, conclusions, or recommendations.

Alexios Mantzarlis, *Indicator*

Basile Simon, *Starling Lab*

Brandon Epstein, *Magnet Forensics*

Carrie Winfrey, *Okthanks*

Daniel Appelquist, *Samsung Open Source Group*

Eliana Quiroz, *Fundación Internet Bolivia*

Gabriela Ivens, *Human Rights Watch*

Gus Hosein, *Privacy International*

Harlo Holmes, *Freedom of the Press Foundation*

Heather Marie Leson, *The International Federation of Red Cross and Red Crescent Societies*

Hoda Hamouda, *PhD, University of British Columbia*

Ila Schoop Rutten, *Communicating with Disaster-Affected Communities Network*

Ingo Boltz, *The Carter Center*

Jhanvi Anam, *Internet Freedom Foundation*

John Reichertz, *Independent journalist*

Mallory Knodel, *Social Web Foundation*

Mario Peña, *SafeCreative*

Martín Szyszlican, *Abrimos.Info*

Nathan Freitas, *The Guardian Project*

Olaf Kolkman, *Internet Society*

Richard W. Vorder Bruegge, *ForensicVB*

Seerat Khan, *Digital Rights Foundation*

Tamas Foldesi, *The International Federation of Red Cross and Red Crescent Societies*

Viktor Kewenig, *Microsoft*

Neeltje Berger, *Microsoft*

Additional reviewers contributed anonymously.

Table of Contents

Executive Summary

Verifying the authenticity of digital content is increasingly necessary. Generative AI has made synthetic media cheap to produce and hard to detect, putting the information environment under sustained pressure. Content provenance technology responds by making the origin and history of digital content verifiable: who made it, with what tools, and whether it has been modified. The Coalition for Content Provenance and Authenticity (C2PA) is the leading effort in this space. Backed by most major technology companies, embedded in cameras, editing software, and AI generation tools, and referenced in legislation across multiple jurisdictions, the C2PA is on a path to becoming foundational infrastructure for how digital content is created, distributed, and verified online.

Provenance technology is a necessary part of restoring trust in the information environment. As it becomes more pervasive and embedded in regulation, it also needs to be scrutinized. Surveillance misuse, if normalized or unaddressed, will undermine the trust that makes content provenance valuable in the first place. That gives the C2PA community, and the governments and platforms deploying its infrastructure, a direct interest in getting this right.

This report maps seven pathways through which content provenance infrastructure can be turned into a tool for identity disclosure, behavioral profiling, and expression control. These pathways do not require the technology to malfunction. They require only that it be implemented in adverse political and regulatory contexts — and the C2PA specifications, built to serve an enormous range of legitimate uses, cannot by itself prevent that. The gap is in the governance ecosystem surrounding it.

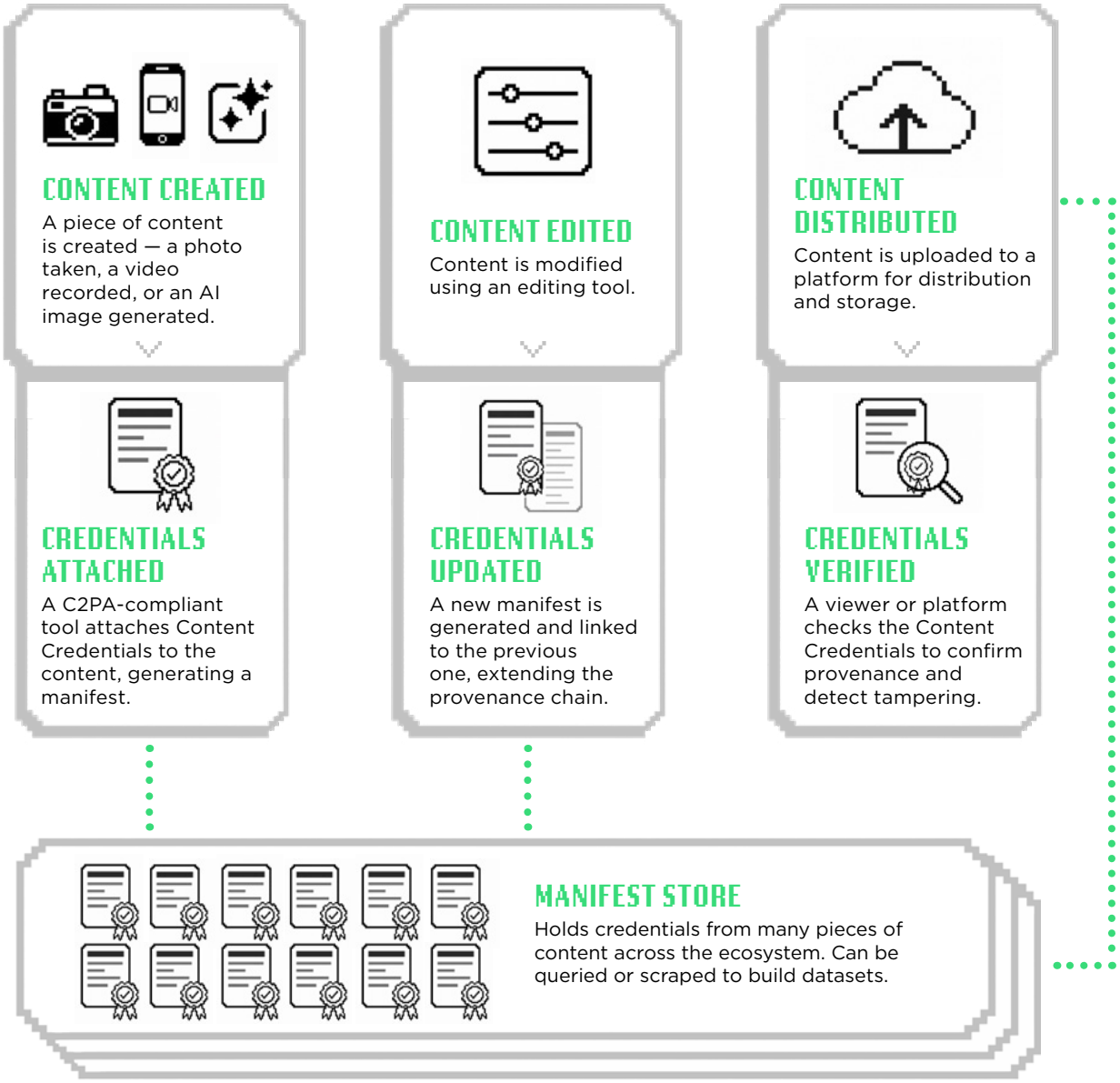
A distinction in the architecture is worth highlighting in this analysis. The C2PA specifications do not allow direct capture of identity information. Add-on technologies, most notably the Creator Assertions Working Group extension, can layer it on top, but the base specifications keep the two separate. This matters because it lets an actor interested only in the "how" of an asset, the tools and edits behind it, document that without identity being bound to the content by default. That the C2PA has kept it separate from the core specifications is a deliberate governance choice, and one worth safeguarding as pressure to converge the two increases.

The populations most exposed are journalists, human rights defenders, and documentary filmmakers. For these groups, content provenance infrastructure creates a distinct and underappreciated surveillance surface: one that links identity to specific digital content with cryptographic precision, accumulates into detailed behavioral profiles over time, and is made harder to contest by the regulatory legitimacy surrounding it. Viewers of credentialed content face their own exposure: the act of verifying content can generate a behavioral record without their knowledge or consent.

The governance gap this report documents is a present vulnerability. The pathways described here are already plausible or occurring in nascent form. The C2PA governance structure currently has no mechanism to assess whether a specific deployment constitutes misuse, and no authority to respond when it does. That gap needs to be filled before deployment norms and regulatory frameworks mature.

These risks are not exclusive to the C2PA, nor to content provenance infrastructure as a category. They are broader questions about how identity, behavior, and expression are governed at scale. Understanding how the C2PA's specific architecture and adoption trajectory shape, amplify, or redirect these risks is what makes it a useful case for examining them.

PRIMER:
What are C2PA Content Credentials?



The diagram above shows how Content Credentials attach to content at creation, accumulate across edits, and flow into manifest stores that hold provenance records from across the ecosystem — available for viewers and platforms to check when deciding whether to trust what they are seeing.

The C2PA does not directly capture identity information. Extension technologies are required for this.

C2PA Content Credentials are a verifiable record attached to a digital file —an image, a video, an audio clip— that describes where it came from, how it was made, and whether it has been changed. Like a recipe that travels with the dish, they carry a readable history of every ingredient used and every step taken, available to anyone who wants to understand what they are looking at [1].

The system behind it is built by the Coalition for Content Provenance and Authenticity (C2PA), an open technical standard backed by most major technology companies, including Adobe, Google, Microsoft, Meta, Sony, and others [2]. Content Credentials are already embedded in devices and tools many creators use today — and increasingly visible to the audiences who encounter their work.

The C2PA is rapidly becoming foundational infrastructure for how digital content is authenticated online. Legislation in multiple jurisdictions is beginning to reference or require provenance standards. Platforms are integrating Content Credentials into publishing and moderation workflows. Camera manufacturers are embedding signing capabilities directly into hardware [3]. For journalists, human rights defenders, documentary filmmakers, activists, and anyone who creates or consumes digital content in high-stakes contexts, this infrastructure will increasingly shape what can be published, verified, and trusted — and by whom.

Participation in the ecosystem is governed by the C2PA Conformance Program, which the C2PA designs, administers, and enforces [4]. Tools, platforms, and certificate authorities that want their Content Credentials to be recognized as valid must meet defined technical and operational requirements and appear on the C2PA's conforming products list. The Conformance Program is the mechanism through which the C2PA shapes who can participate in the ecosystem and on what terms.

The C2PA Today: Governance, Implementation & Regulation

The information presented in this section reflects the state of the ecosystem at the time of publication. The governance of the C2PA, its adoption and regulation are moving quickly .

THE DESIGN AND GOVERNANCE OF THE C2PA SPECIFICATIONS AND ECOSYSTEM

The C2PA is governed by a Steering Committee of member organizations that set the strategic direction of the standard, approve specification releases, and oversee the Conformance Program that determines which tools and platforms are recognized as part of the ecosystem [5]. It is made up of:

Adobe, amazon, BBC, Google, Meta, Microsoft, OpenAI, PUBLICIS GROUPE, SONY, truepic

The C2PA Technical Working Group develops the specifications and its accompanying documentation, including the Harm Assessment [6], UX Recommendations [7], and Guidance for AI and Machine Learning [8]. Participation is open to all C2PA members. Civil society engagement remains extremely limited relative to the influence these processes carry. Organizations working on press freedom, human rights documentation, and digital rights have a direct stake in how the specification evolves — and a standing invitation to get involved.

For more information on membership and participation, visit c2pa.org/membership.

C2PA IN THE WILD

The C2PA is already embedded across the entire content lifecycle — from the moment content is captured to the moment it is verified by an audience. The examples below are selective; the full list of conforming products is available at the C2PA Conformance Explorer.

Capture

Hardware with C2PA signing built into the device.

- Leica M11-P
- Sony Alpha 9 III
- Nikon Z6 III
- Canon EOS R1

Mobile & OS

Signing at the platform layer, before any third-party app.

- Google Pixel
- Samsung Galaxy S25
- Xiaomi Camera
- Qualcomm Snapdragon

Edit & Produce

Software that maintains the provenance chain through post-production.

- Photoshop
- Lightroom
- truepic
- PIXEL STUDIO

Generate

AI tools that attach Content Credentials at the point of creation.

- Adobe Firefly
- Microsoft Bing Image Creator
- stability.ai
- OpenAI

Distribute & Verify

Platforms and tools that read, display, or require Content Credentials.

- LinkedIn
- YouTube
- Instagram
- TikTok
- content credentials verify

Selected examples only. Full list at the C2PA Conformance Explorer: contentcredentials.org/verify.

REGULATORY LANDSCAPE

The regulatory landscape around content provenance is evolving rapidly and across jurisdictions. The frameworks below vary significantly in scope, intent, and rights protections: some are designed primarily around consumer transparency, others around platform accountability, and at least one is designed in ways that may activate one of the surveillance pathways this report documents. None were developed with meaningful civil society participation, though the EU AI Act Code of Practice is a partial exception: it includes a formal civil society track, and engagement in that process remains open and consequential. For all other frameworks listed here, civil society was largely absent from the drafting process. Most of these frameworks do not name the C2PA explicitly, and are more accurately described as technology-agnostic provenance or labeling mandates that the C2PA is positioned to satisfy. The examples below are selective and reflect the situation at the time of publication [9-17].

Jurisdiction & Instrument	Status	Requirements	Rights/civil society notes
EU EU AI Act, Article 50 + Code of Practice	Article 50 in force Aug 2026; CoP in development	Machine-readable marking of AI-generated content & disclosure; CoP to operationalize provenance specifications	Most developed rights protections; civil society participation formally included in CoP; gaps remain on viewer privacy and open-weight models.
California SB 942 (2024) + AB 853 (2025); AB 2713 + SB 1000 pending	SB 942 + AB 853 enacted; operative Aug 2026; amendments active	Machine-readable marking of AI-generated content & disclosure; CoP to operationalize provenance specifications	Personal provenance data definition is a significant privacy provision; no formal civil society engagement mechanism; amendments in progress.
China Measures for Labeling AI-Generated Content	In force Sept 2025	Explicit and implicit labels on all AI-generated content; embedded metadata must include service provider identity and unique content ID	This framework operates independently of the C2PA. It requires China's own mandatory national metadata standard. It illustrates the same underlying risk through a different vehicle: state-mandated provider identity embedded in content metadata, which produces conditions comparable to the architectural capture pathway described in this report without the C2PA being involved at all. User protections are included, but no independent enforcement mechanism is established. Whether the framework activates surveillance risks in practice depends significantly on implementation context and enforcement.
India IT Intermediary Rules Amendment, G.S.R. 120(E)	In force Feb 2026	Mandatory labeling and provenance for synthetically generated information; platform proactive detection obligations	The rules place compliance obligations on intermediaries without corresponding requirements on AI developers at the point of creation. Provenance metadata is required but no open standard is referenced, meaning provenance data will not travel across platforms or borders. Liability for wrongful removal is absent, creating over-removal incentives. Civil society feedback was submitted during drafting; it is not known whether or how it was considered in the final text. Enforcement framework not yet tested.

Six Ways the C2PA Infrastructure Can Expose Your Personal Information

The exposure types described below are not the result of flaws in the C2PA specifications. The standard is built to serve an enormous range of legitimate uses — consumer cameras, professional broadcast workflows, AI-generated content platforms, news wire authentication, and many more. A specification narrow enough to prevent every possible misuse would be too constrained to function as shared infrastructure. That breadth is a deliberate and reasonable architectural choice.

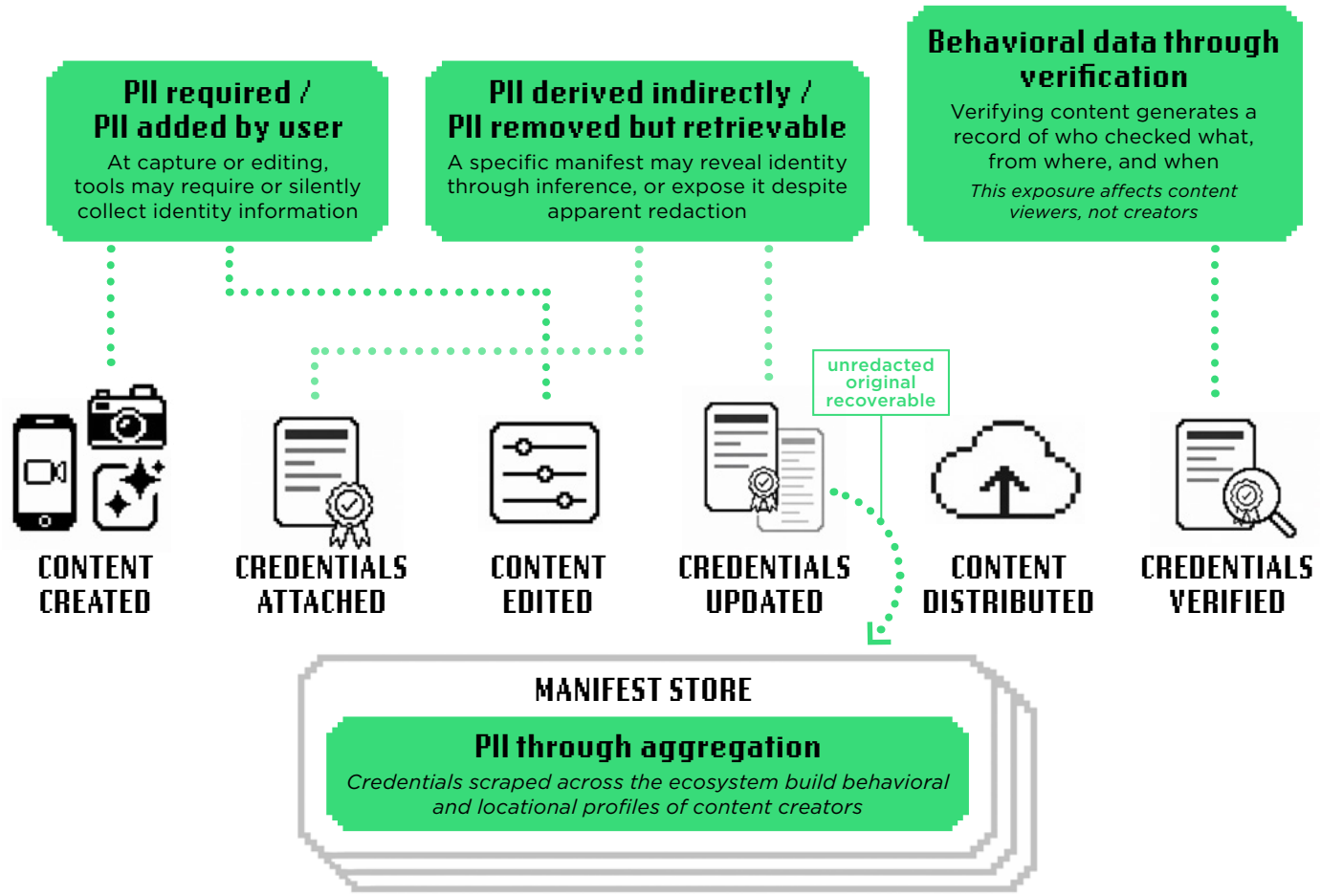
It also means that some misuses cannot be addressed at the specifications layer. Whether the standard is deployed safely falls instead to the contexts in which it operates: the laws that mandate it, the governance structures that oversee it, and the tools that implement it. Three activation pathways are worth naming before the exposure types, because they shape what kind of response each requires.

The first is legislative and regulatory misuse. A government that understands the C2PA's privacy surface can exploit it deliberately — through mandated identity assertions, required credentials as a condition of distribution, or convergence with national identity systems. The more likely near-term risk, however, may be a well-intentioned regulator who mandates C2PA-compliant credentials without understanding what that mandate activates. The outcome can be functionally identical to deliberate misuse.

The second is architectural capture. A state or hostile actor does not need to pass a law to weaponize C2PA infrastructure. By requiring creators to obtain identity credentials from a state-controlled authority as a condition of producing or distributing credentialed content, it can build two compounding functions into the infrastructure by design: a gatekeeping mechanism that controls who may produce recognized credentialed content, and a surveillance registry that links every credentialed creator to every piece of content they have produced — all without any requirement for legal process.

The third is design and implementation failure. Platforms and tool operators can extract surveillance value from validation infrastructure or certificate data without any adversarial intent. Tool designers may not adequately surface privacy implications to users. No hostile actor is required. The harm assembles itself from the gap between what tools do and what users understand them to do.

Each pathway is illustrated in the stories in the section that follows.



Identity can be required as a condition of creating or distributing content. A law or platform policy may require attaching personal information to Content Credentials before content can be published or distributed. The C2PA specification does not prohibit this as mandatory identity assertions may, in specific use cases, be a legitimate use of the standard. A government mandate requiring journalists to register their identity with a national authority before their content can carry verified credentials would require no modification to the specifications whatsoever, and would not be distinguishable, at the infrastructure layer, from those legitimate uses[1].

Illustrated in Scenarios 1 and 2

Personally identifiable information can be added by the user — inadvertently, or without being informed of the privacy implications of doing so. Content Credentials can carry personal information added by the creator—a name, a caption, a device identifier—without the tool surfacing what that disclosure means or who can access it. The harm is not always intentional on the part of the platform: tool design that prioritizes functionality over privacy literacy can produce the same outcome as deliberate data collection. A photographer including personal attribution to an image may not realize that information will travel permanently with the file, accessible to anyone who inspects the manifest.

Illustrated in Scenario 7

1 This would require implementing the extension technology of the Creator Assertions Working Group (CAWG) onto the C2PA [18]

Identity can be inferred from data that appears anonymous. Assertions and certificate information within a manifest may appear innocuous individually, but can be cross-referenced with external registries or issuance records to deanonymize. Location data, device identifiers, workflow metadata, a certificate serial number — none of these need to be sensitive in isolation, but when combined, or matched against external records, they can narrow an anonymity set to a single individual.

Illustrated in Scenarios 3, 5 and 6

Redacted identity information may remain recoverable. Although the C2PA allows for redaction of fields within Content Credentials, soft binding techniques such as watermarks or fingerprints, a feature of Durable Content Credentials, may be used to locate the original, unredacted manifest [19, 20]. For instance, a human rights documenter who redacts their name from a manifest before publishing, unaware that a watermark embedded in the image still points to the original signed file.

Illustrated in Scenarios 6 and 7

Identity can emerge from patterns across a body of published work. Identity may become recoverable not from an individual manifest but from correlating assertions across a body of work over time—locations, timestamps, device identifiers, behavioral signatures—none of which individually crosses a sensitivity threshold, but which together build a detailed profile. For example, a state actor scraping a manifest store to map the movement patterns of an activist photographer across months of published work would not need access to any single sensitive file.

Illustrated in Scenarios 3, 5 and 6

Engaging with Content Credentials exposes creator and audience behavior to third parties. Engaging with Content Credentials — whether as a creator signing content or as an audience member verifying it — can expose behavior to third parties. On the creation side, signing operations that require external connections for timestamping, certificate status checks, or manifest store submission generate server-side records linking the creator's device, location, and timestamp to a specific piece of content, without any disclosure that this is occurring. On the verification side, depending on implementation, remote validation may require the viewer's device to contact an external server directly, generating a logged request that records who verified what, from where, and when. In neither case does the affected party have awareness that this is happening or any means of refusing it: unlike cookies or tracking pixels, the C2PA specifications include no consent mechanism, no opt-out, and no disclosure requirements. A journalist signing footage before publication may unknowingly leave a server-side trace of that act. A reader who encounters a suspicious image on social media and verifies its provenance may unknowingly send a request associating their IP address, approximate location, and timestamp with that specific piece of content. At scale, across a platform or a jurisdiction, these logs become a map of who is creating what and who is reading what, where and when.

Illustrated in Scenario 4

Stories of Content Surveillance

The research underpinning this report maps seven distinct scenarios through which content provenance infrastructure can become a tool for surveillance, identity disclosure, and expression control.

The scenarios are based on adversarial modeling: a structured method for identifying how a system can be turned against the people it is meant to serve. They were developed from direct knowledge of the C2PA specification, the governance structures surrounding it, and the legislative frameworks referencing it across multiple jurisdictions. Each is intentionally specific — named characters, plausible regulatory instruments, concrete sequences of events. That specificity is not a claim that harm will occur in exactly these forms. It is a method for making harm pathways visible, which is harder to do in the abstract. Each scenario is grounded in documented real-world precedents: laws that exist, technical systems already deployed, enforcement mechanisms that have been used in analogous contexts.

Following an initial draft, the scenarios were reviewed by a range of experts across journalism, human rights, digital security, and standards development. The ask was not for validation but for scrutiny: where do the scenarios hold, where do they break, and where do they understate or overstate the risk. The scenarios in this report reflect that process.

Story One:

The Chilling Effect

Amara, freelance journalist

Amara is a freelance audiovisual journalist based in a mid-sized city. She covers local politics, labor disputes, and the occasional cultural story. She films, edits, and submits short video reports and photo essays to the three or four outlets she works with regularly. She is not famous. She is not on any watchlist. She owns a mid-range camera and uses a popular editing application to process her work. Both are C2PA-enabled.

When her government passes a Media Integrity Decree — the hypothetical regulatory instrument in this scenario — framed as an anti-disinformation measure, it requires all news content distributed via licensed platforms to carry valid Content Credentials. The credentials must include a state-issued press credential number linked to the journalist's registration record. Only certificates issued to state-accredited journalists and outlets are accepted on the state trust list. Platforms implement this as a distribution gate: content signed with certificates not on the list is flagged as unverified and deprioritized; content without credentials is treated as presumptively suspicious.

Amara registers. The process is bureaucratic but not impossible. She gets her license number, configures her tools, and continues working. The first few months are uneventful. Her content carries her credential. Editors can verify its origin. She considers this, cautiously, a reasonable tradeoff.

In the autumn, she begins filming a story that is more sensitive: a series of government contracts awarded to companies with documented connections to senior officials. The footage is solid, and her editor is interested, but Amara hesitates before submitting.

Publishing means attaching her credential — her name, her registration number, her verified identity — to a piece of journalism that directly implicates people in power. She is not wrong to hesitate. She has seen what happens to journalists who make powerful people uncomfortable. Her credential, which felt like a reasonable tradeoff in the spring, now feels like a liability.

She makes a decision: she will not publish under her name. She strips the personal information from the manifest using the redaction function in her editing software and publishes the footage anonymously through a smaller, unlicensed platform that does not require credential verification. She cannot use the outlet she normally works with. That outlet operates under a platform distribution license that requires credentials on the state trust list. Publishing without her credential there is not an option.

The economic consequences are immediate. The outlet that wanted the story no longer has it. The relationship with that editor — built over years of reliable work — is strained. The fees she would have earned are gone. Content published on unlicensed platforms is easier to dismiss: the government's communications apparatus is practiced at labeling inconvenient journalism as unverified or fabricated. Without her credential and without the institutional backing of a licensed outlet, the footage is more vulnerable to exactly that framing. She publishes anyway, because the alternative is not publishing at all.

She thinks that is the end of it, but what Amara does not know is that the manifest, stripped of her name, still carries the fingerprint of her working practice: the device identifier from her camera, the editing software signature, the specific sequence of adjustments she applies, with minor variations, to almost everything she shoots. None of these fields contain her name. Each of them, individually, means little. Together, and compared against the months of credentialed work she has published under her name, they form a pattern that is functionally a signature. The mechanism that made this possible is developed in full in Scenario 6, The Trail.

Story Two:

The Compromising Bridge

Marcelo, anonymous creator, satirist

For three years, a series of short animated clips has been circulating on social media: sharp, funny, and precise in the way that only someone with genuine inside knowledge can be precise. The videos mock government officials by name, reconstruct leaked procurement documents as puppet theatre, and have on several occasions, broken stories that the mainstream press picked up days later. The creator makes an effort to remain anonymous.

The person behind this account is Marcelo. He is thirty-one years old, works a day job in graphic design, and has been making the videos on evenings and weekends since a particularly egregious corruption scandal convinced him that someone needed to say what the official press would not. He is careful. He uses a separate device for the videos, a VPN, anonymous accounts, and cash payments for the software he needs. He has thought seriously about operational security.

He is also, like most visual content creators, increasingly reliant on AI-assisted tools. His current editing software uses a generative AI layer for background rendering and motion graphics since it saves him hours of work and makes the videos look more professional. He upgraded to it eight months ago.

At the time of the upgrade, his government had recently passed an age verification law requiring platforms offering AI generation tools to verify that users were adults before granting access. The law was framed as child protection. It passed with broad public support and minimal opposition. Marcelo verified his age using his national ID — the same credential he uses for banking, tax, and healthcare. It felt identical to a hundred other routine identity checks he had completed in other contexts. He did not think about it again.

What Marcelo did not know was that eight months later, a separate piece of legislation (focused on tackling disinformation rather than child protection) would require that all AI-assisted content distributed via registered platforms carry valid Content Credentials, including an identity assertion authenticated against the national digital identity system. The platform already has his government ID number from the age verification process. The new regulation instructs it to attach that identity to the Content Credentials of all AI-assisted content produced by verified users. No new action is required from Marcelo. No prompt appears. The updated terms of service are updated, but Marcelo accepts them without reading. The bridge between the age verification infrastructure and the content provenance infrastructure is made in the regulatory and platform backend.

Marcelo continues making videos. He believes he is anonymous. He has no reason to think otherwise. He has not changed anything about how he works, has not clicked through anything new, has not made any conscious decision he would recognize as a security risk.

One evening, several months later, he is reviewing the metadata of a recently published video. He opens the manifest. He reads through the fields he expects to see: the software signature, the timestamp, the rendering parameters. And then, in a field he does not immediately recognize, he sees a number. It takes him a moment to realize that it is his national ID number.

Story Three:

The Registry

Lorena, Editor

The outlet has been running for over a decade. It started as a blog covering local government and grew into a small but credible newsroom. It is not famous outside its city, but readers trust it in the way that people trust something they have watched to be consistently right, or at least professional, over a long time.

Lorena has been its editor for four of those years. The outlet has been under pressure for about two years — no direct intimidation, no shutdown orders, but enough to make her careful about which stories carried the outlet's name prominently and which did not.

When a foreign investigative outlet approached her team to collaborate on a regional investigation, the structure they agreed on was straightforward and common practice: the local outlet would contribute reporting, footage, and processed images; the published story would carry only the foreign outlet's byline. The local outlet's involvement would not be confirmed or denied. Cross-border collaborations structured this way are standard risk management as they distribute exposure, and the foreign partner's international visibility provides a layer of protection the local outlet cannot provide for itself.

Lorena's team processed the contributed content on their standard equipment: C2PA-conformant cameras and an editing suite. When they had purchased the equipment the previous year, they had filed the required registration with the press authority. The form included a field requesting the signing certificate identifier for each registered device. The ministry had described it as a technical update for media monitoring purposes. Lorena had filled it in without particular concern.

The story was then published under the foreign outlet's byline. Within two weeks, a source told Lorena that a ministry official had asked, in an unrelated meeting, whether her outlet had been involved. The question was casual but also too specific.

She found out later how the connection had been made. The certificate identifier was publicly readable in the metadata of every piece of C2PA-signed content her team had contributed. The ministry had queried the device registry. The registry returned the outlet's registration record. The connection between the content and the outlet had been confirmed without interception, without a court order, without approaching the foreign outlet, and without any legal process. The form filed as a routine compliance step had been the bridge.

The ministry had not yet acted on the information formally. But Lorena understood what it meant to have that confirmation in a government database. The protection the collaboration structure had been designed to provide had been quietly removed, at the moment the equipment registration form was submitted, before the investigation had even begun.

Story Four:

Watched While Watching

Sofia, hospital administrator

Sofia is thirty-four years old. She works in hospital administration, has lived in the same city her whole life, and follows the news more carefully than most people she knows. She is not an activist. She votes, she reads, and she has become, over the past few years, increasingly careful about what she believes online. She fact-checks things before she shares them. She has started using content verification tools when something seems off — a video that looks real but feels staged, an image circulating too fast, a clip of a politician that doesn't quite match what she remembers them saying. She considers this a civic act.

The information environment she navigates has become increasingly polarized. The government has, over several years, developed an informal but consistent position on which sources are trustworthy and which are not. International human rights organizations are characterized as foreign-funded interference. Multilateral bodies such as UN agencies or regional human rights courts are described as ideologically captured and anti-national. Several domestic opposition-aligned outlets have been labeled as destabilizing. None of this is law, exactly, but it is part of an ominous atmosphere — the sort of thing that shapes how institutions think without being written down anywhere that could be challenged in court.

Sofia does not particularly agree with this framing. She reads what she reads because she is trying to understand what is true. When content from these sources circulates — a report on detention conditions, a dataset on electoral irregularities, a video of a demonstration — she verifies it. Not because she endorses it, but because she wants to know if it is real.

What Sofia does not know is that every verification request she makes is a network event. Her app does not validate content locally; it contacts a remote server to check the content's manifest against a repository. That request carries information: what content she is checking, when, from which device, and from which location. The platform that operates the verification infrastructure logs these requests. This is standard practice. The logs are used for analytics, performance monitoring, and, as the terms of service note, improvement of the service and compliance with applicable law.

Over eighteen months, Sofia has verified content from international human rights organizations, UN agency reports, regional human rights court rulings, and domestic outlets that the government has publicly described as opposition-aligned. She has also verified content from government sources, and from other mainstream outlets. But the pattern is there, legible to anyone looking: a sustained and recurring attention to sources that the government has characterized as ideologically suspect.

She is not the only one. Millions of people use the same verification tool.

In the autumn, Sofia applies for a position in the public health administration. It is for a senior role she is qualified for, has been working toward for three years, and has every reason to expect she will get. The application process is standard: qualifications review, interviews, reference checks. There is also, as there is for all government positions above a certain grade, a background and suitability assessment. The criteria for that assessment are not published in full.

Her application is rejected. The rejection letter cites insufficient suitability for the role. No further explanation is provided. She is told she may reapply after two years.

Sofia does not know why. She reviews everything she can think of. Her work record, her references, her interview. She cannot find the reason. She asks a contact in the ministry if there is any informal feedback available. There is not.

What she cannot see is that her verification history –logged, retained, and accessible to government agencies under the platform's compliance obligations– was reviewed as part of her suitability assessment. The pattern of her attention to ideologically suspect sources was noted. No individual verification event was decisive. The aggregate was.

Story Five:

The Certificate Fingerprint

Tariq, humanitarian worker

The organization has been at the forefront of documenting atrocities for years. It trains local partners, develops documentation protocols, and maintains a small field presence in active conflict zones. Its footage has been submitted to international tribunals, cited in UN reports, and used in asylum proceedings. The credibility of that footage is critical.

Two years ago, the organization implemented C2PA Content Credentials across its field documentation work. The decision was deliberate and considered. Fabricated footage was circulating in the same conflict zones where the organization worked, undermining the credibility of genuine documentation. But the organization was also acutely aware of the safety risks that identity disclosure creates for field staff. So their implementation was carefully scoped: Content Credentials would establish the how of their footage, not the who. Device-level signing would confirm that content was captured on an unmodified device, unaltered since capture, at a specific time and place. No organizational identity. No staff identifiers. The credential would prove the integrity of the content without revealing who produced it.

It was, by any reasonable standard, a privacy-conscious implementation. The organization had thought carefully about the tradeoff and made a considered choice.

Tariq is one of three field staff the organization rotates through the region. He has been with the organization for six years, and has contacts in communities that no one else on the team can reach. He is careful. He varies his routes and does not carry identification connecting him to the organization when he is in the field.

In the spring, he documents a series of incidents in a district that has been the site of sustained military operations. The footage is significant, showing destruction of civilian infrastructure in a pattern consistent with the targeting allegations the organization has been investigating. He captures eleven clips over four days. Each is automatically signed at the moment of capture using the organization's documentation kit: a ruggedized camera and a mobile application, both C2PA-conformant.

He submits the footage through the organization's secure channel. It is processed, verified, and prepared for submission to an international monitoring body.

What neither Tariq nor the organization has fully accounted for is what the credential record reveals through a different pathway entirely. The device-level signing does not disclose the organization's identity. But it does disclose the tool. And the tool is the problem.

The documentation kit the organization uses is C2PA-conformant, but it is not widely adopted. Globally, it has a few thousand users. In this conflict zone, in this period, the number of actors using this specific combination of hardware and software is very small. The tool signature embedded in every credential record is not a name, but in this context, it functions like one.

The state actor does not need a surveillance program to make the connection. They need two things that are already publicly available. The first is the organization's own archive. In regions where field staff safety is less of a concern, the organization signs its content with its organizational identity. It is standard practice, and a source of institutional credibility with the tribunals and monitoring bodies it works with. That archive is public, verifiable, and searchable. It establishes, unambiguously, that this organization uses this specific tool. The association between the tool signature and the organization's name is not inferred. It is proven, repeatedly, by the organization's own publishing practice in contexts where they had no reason to hide it.

The second is the content credential metadata ecosystem. Services that index C2PA manifests, aggregating records from published content across platforms, make the tool signature searchable across a body of work. The conflict zone footage, submitted to the monitoring body and entering a semi-public record, carries the same tool signature as dozens of other pieces of content the organization has published under its name elsewhere.

The tool signature in the conflict zone footage matches the tool signature in the organization's public archive. The organization's known field presence does the rest. The credential record the organization designed to protect its staff contains, in the tool signature alone, a thread that leads directly back to them, and they placed that thread in the public record themselves, in good faith, in a different context entirely. The anonymity set was the user base of that tool, in that region, in that period, and that number was small enough to matter.

The credential record the organization designed to protect its staff, by proving the integrity of footage without revealing identity, turned out to contain a fingerprint they had not anticipated.

Story Six:

The Trail

Amara, freelance journalist (contd.)

This scenario is a continuation of Scenario 1

She stripped the metadata and published anonymously. She thought that was the end of it.

Amara had made her decision carefully. The story she had built was solid: government contracts, documented connections, evidence she had gathered over three months. Publishing under her name was not an option she was willing to take. So she used the redaction function in her editing software, removed any PII she could see, and uploaded the footage to a smaller independent platform that did not require credential verification. She had done what she knew how to do and believed it was enough.

What she could not see and anticipate was the provenance chain.

C2PA credentials do not only record the final state of a piece of content. They can record its history, including the sequence of devices, applications, and editing steps that produced it. Each step in that chain carries a signature. The camera that captured the original footage. The editing application that processed it. The export settings that produced the final file. These signatures are not personal identifiers in any simple sense. They do not say her name, but they do describe, precisely and consistently, the technical environment in which she works, and how she does it.

Over the previous eight months, Amara had published dozens of pieces of credentialed content under her name. Each carried the same chain: the same camera model with the same device certificate, the same editing application with the same software signature, the same sequence of adjustments she applies to almost everything she produces. That body of work —published openly, signed with her identity, accumulated through her compliance with the Media Integrity Decree— constitutes a detailed behavioral baseline.

The anonymous footage carries the same chain. It is not identical; she had changed some settings, and even used a different export format... but it was similar enough. The camera certificate matches. The editing software signature matches. The sequence of color grading adjustments, the audio normalization pattern, the crop ratio she favors: all consistent with the baseline her credentialed archive established.

This is not a coincidence. It is the provenance chain working exactly as designed. It is designed to record the technical history of content with precision and consistency, so that anyone who looks can understand how it was made. The same feature that makes her credentialed journalism verifiable makes her anonymous journalism traceable.

The comparison does not require sophisticated technology. It requires access to two things: her public credentialed archive, which is openly available, and the anonymous footage, which is now in a semi-public record having been submitted to the platform. A government agency monitoring independent platforms for content critical of the administration has both. The analysis is not complex. The provenance chains are compared. The signatures match within a threshold that makes coincidence implausible.

This is not only Amara's problem. Every journalist, documenter, and content creator who has complied with content credential mandates has been building a behavioral baseline: a detailed, timestamped, publicly accessible record of how they work, accumulated across every piece of content they have published under their name.

The population of people who have complied longest, most consistently, and most completely is the population most committed to working within the system. They are also, now, the most vulnerable. Their compliance with a transparency requirement has made them transparent in a way they did not intend and cannot undo.

Story Seven:

Unwitting & Unwilling Exposure

Alice & Joe

****** Alice ******

Alice is a documentary filmmaker. She spent the last year making a film about informal labor conditions in her city's logistics sector: warehouses, delivery networks, the people who keep things moving. It is careful, human work, and she is good at it.

Her production software is C2PA-enabled. She uses it because her international distribution partners require it. When she installed it, the setup asked for her name, email, and country. Standard fields. She completed them and started working.

What the setup process did not explain is that the software's default configuration attaches her account details to the Content Credentials of every file she exports, via the CAWG identity extension. The option to disable this exists, in an advanced settings panel she has never opened, described in language that assumes familiarity with the C2PA specifications.

For most of the year this does not matter. Then, in the final weeks of production, she films something unplanned: a confrontation between managers and workers organizing without official recognition. She decides to submit the clip anonymously to a press freedom organization abroad. She exports it without checking the Content Credentials panel, because she does not know there is anything there that needs checking.

Her name travels with the file. When she tries to understand what happened and what recourse she has, she finds that the software's support documentation is in English only, the privacy policy is governed by US state law, and the process for requesting credential correction routes through a legal mechanism she has no practical access to. She submits a support ticket. She does not receive a reply.

****** Joe ******

Joe covers conflict. He is a photojournalist filing for international outlets from places they rarely send staff. He has taken security training. He knows how to strip EXIF data.

He also wants attribution. Photographs taken in difficult places have a habit of circulating without credit. When he learned his C2PA-enabled tools could embed verified authorship credentials that travel with the image, he opted in deliberately: opened the settings, connected his account, enabled the CAWG identity extension.

For eight months this works exactly as intended. Then, he is in a location he cannot name publicly, documenting something that could put him in danger if his presence were known. He photographs, edits, exports. He asks his editor to publish without his byline, citing security concerns. The editor removes his name from the caption. Neither of them checks the credential.

The credential still carries his name. He enabled it eight months ago, in a different country, under different circumstances, and the interface gave him no indication that anything had changed. The setting ran quietly in the background of every file he had produced since.

It was his choice. He made it in good faith. The circumstances changed. The tool did not tell him.

What Makes C2PA-Enabled Surveillance Different

The risks described in the preceding stories are not simply new instances of familiar surveillance problems. They share structural properties that make them harder to detect, harder to contest, and harder to remedy than other existing forms of digital tracking. Understanding those properties matters because the safeguards designed to address platform behavioral data or communications metadata were built with a different threat model in mind. They do not map cleanly onto what the C2PA ecosystem makes possible.

Five properties are worth naming clearly.

Content-specific tracking

While a search query can reveal that someone is interested in a subject, or a social media follow an affiliation, a C2PA event — whether a signing event by a creator or a verification event by a viewer — reveals something more precise. For creators: that a specific person produced a specific, cryptographically identified piece of content, with a specific tool, from a specific location, at a specific time. For viewers: that a specific person engaged with a specific, attributable piece of content, at a specific moment, from a specific location. In both cases, that precision is what makes C2PA data more actionable for profiling purposes than general browsing or publishing behavior. It does not just show what someone is interested in or working on. It links identity to specific expressive and epistemic acts with cryptographic certainty.

The epistemically careful are the most exposed

General browsing surveillance captures everyone indiscriminately. C2PA surveillance — whether of signing or verification behavior — captures specifically the people most likely to be doing accountability work, source protection work, or opposition research: journalists, researchers, human rights monitors, documenters, lawyers, and engaged citizens. Creators who sign content with C2PA credentials are self-identifying as people who produce credentialed documentation. Viewers who verify content are self-identifying as people who are trying hardest to establish what is true. Both populations overlap substantially with those of greatest interest to governments with authoritarian tendencies. Surveilling either adds a high-signal layer that general behavioral surveillance does not produce.

The trust gap

Most people understand, at some level, that browsing is tracked. C2PA is framed explicitly as a trust and transparency tool — something that protects users from manipulation and makes the information environment safer. Users who verify content are making an active choice to engage with a trust infrastructure. That framing is not false, but it actively suppresses awareness of the surveillance surface the tool creates. The gap between the tool's stated purpose and its surveillance potential is wider, and more consequential, than for ordinary browsing because the

tool's identity as a trust mechanism is precisely what makes users least likely to approach it with caution. A person checking content authenticity is not expecting that act to generate a behavioral record. That expectation gap is structural, not incidental, and it cannot be closed by better privacy notices alone.

Standardization as a force multiplier

Existing behavioral surveillance is fragmented across platforms, each with its own data architecture, retention policy, and technical format. The C2PA creates a standardized, interoperable, machine-readable infrastructure that produces consistent structured data across platforms, tools, and implementations — dramatically lowering the cost and increasing the reliability of surveillance that would otherwise require significant resources to conduct. Critically, the C2PA is not a feature of any single device or platform. It is becoming embedded across the entire content ecosystem, from cameras and editing tools to AI generators and social media platforms. The result is not a surveillance capability attached to one tool, but one woven into the infrastructure through which digital content is created, edited, distributed, and verified at scale. Advances in large-scale AI inference compound this: the accumulated, structured record of who created what, with what tool, from where, when, and who sought to verify it — generated as a routine byproduct of the infrastructure functioning as intended — can now be processed into actionable intelligence at a speed and scale that earlier capability thresholds would not have permitted.

Regulatory legitimacy as cover

Governments that compel access to general browsing data face legal friction and civil society opposition. That demand is clearly identified as surveillance and contested as such. Governments that compel access to content verification logs under platform accountability or AI transparency legislation have a structurally cleaner narrative: they are enforcing transparency requirements that exist for the public good. The regulatory framing that makes the C2PA politically viable — as a response to disinformation, as an AI transparency measure, or as a tool for information integrity — simultaneously makes compelled access to its infrastructure easier to justify, harder to oppose, and less visible as a civil liberties concern. The mechanism that makes the C2PA valuable to its intended users is the same mechanism that makes state access to its data appear legitimate. Contesting it requires a vocabulary for identifying C2PA-enabled surveillance that is legible outside the standards community — to courts, legislatures, and press freedom organizations — and that vocabulary does not yet exist.

A Roadmap for Prevention: Governance and Red Lines

The scenarios in this report demonstrate that the surveillance risks in the C2PA ecosystem cannot be resolved at the technical layer alone. Work is underway within the standards community on privacy-preserving alternatives, including, for example, signing approaches that reduce the linkability of credentials and stronger minimum standards for anonymity. Although that work is necessary, the core finding of this research is that most of the pathways documented here do not require any flaw in the technical standard to activate. There is no specification change that prevents a state from mandating identity assertions as a condition of distribution, and no certificate design that protects an outlet when a government registry bridges publicly readable identifiers to named publishers. The gap and opportunity is therefore in the governance ecosystem surrounding it — which currently has no mechanism to recognize egregious misuse when it occurs, no responsible body with the mandate and independence to assess it, and no defined consequences for actors whose deployment of C2PA infrastructure causes serious harm.

The gap is not only institutional. Journalists, human rights defenders, and documentary filmmakers who rely on this infrastructure daily often do not know what a credential exposes, and viewers verifying content rarely know what that verification records about them. Closing the governance gap without closing this literacy gap leaves the populations most exposed to bear risks they cannot see and did not consent to. Prevention requires both: informed communities who can make deliberate choices about disclosure, and governance mechanisms that hold the ecosystem accountable when those choices are overridden or ignored.

Literacy as a first line of defense

A journalist who signs content without understanding that the credential can be traced back to a device, a location, or a pattern of activity cannot make an informed decision about disclosure. A viewer who verifies a credential without knowing that verification itself can generate a record cannot consent to that exposure, because they are never told it is happening. Addressing literacy gaps such as these are required to protect privacy. Actions in at least two levels are needed to close this gap.

First, community-level risk education. Journalists, human rights defenders, documentary filmmakers, and the organizations that support them need concrete, practical guidance on what content credentials expose and how. This means harm modeling tailored to specific contexts of use, developed with input from the communities most exposed rather than delivered to them. Training materials, verification tool documentation, and onboarding flows for provenance-enabled tools should state plainly what is recorded, who can access it, and what cannot be undone once a content credential is attached to an asset.

Second, media literacy at the point of consumption. Viewers encountering credentialed content need to understand not only how to interpret content credentials, but what interacting with one may reveal about them. Public media literacy efforts, though still insufficient, have focused on teaching audiences to recognize and trust provenance signals. They have not yet addressed the reciprocal risk, that the act of checking a credential can itself be logged, and what alternatives they may have at their disposal to authenticate content without undermining their privacy.

Literacy does not substitute for governance. A well-informed journalist still cannot decline a government mandate to include identity assertions in published content. But literacy determines whether affected communities can recognize misuse when it occurs, advocate for the protections this report recommends, and make deliberate rather than uninformed choices about the infrastructure they now depend on.

Accountability mechanisms for a trust infrastructure

Accountability mechanisms can operate at more than one level, and the right architecture need not be resolved before the case for them is accepted. The most immediate opportunity is within the C2PA ecosystem itself. The C2PA has the standing, the conformance infrastructure, and the trust list authority to establish an arms-length oversight function before external actors fill that space with less accountable alternatives. Taking that step would position the C2PA as a governance leader in the content provenance space, ahead of regulatory intervention. The risks documented here are not unique to the C2PA, however. They attach to content provenance and authenticity infrastructure broadly, and to content transparency mandates more broadly still. An oversight function with a mandate scoped to that wider domain — whether housed within an existing multi-stakeholder body, a press freedom coalition, or a purpose-built institution — would have both the independence and the breadth to address harms that arise across different standards and deployment contexts. The case for C2PA-internal governance and the case for independent external oversight are not mutually exclusive: each strengthens the other.

Whatever form these mechanisms take, three characteristics are necessary for them to be effective.

Independence. They cannot be controlled by industry implementers or state actors. A mechanism whose composition or funding can be shaped by the actors most likely to commit misuse cannot credibly assess or respond to it.

Multi-stakeholder composition. Civil society, press freedom organizations, human rights defenders, technical experts, and legal scholars must have meaningful roles — not advisory roles, but decision-making ones. The populations most exposed to the risks this report documents must have representation in the structures designed to address them.

A defined mandate and real consequences. Findings must be public. Recommendations must be actionable. A mechanism that can only observe and comment is not an accountability mechanism. The C2PA Conformance Program already has the authority to remove products from its conforming products list and to remove certificate authorities from the trust list. Accountability mechanisms must be positioned to activate those consequences where warranted.

Red lines: an example for the C2PA and a model for the field

The conditions below are framed around the C2PA because it is the dominant standard and the one this report examines in depth. They are also an opportunity: by defining and enforcing red lines within its own ecosystem, the C2PA can take the lead in addressing the surveillance risks that content provenance infrastructure carries — establishing a model for how other standards bodies, platforms, and regulators can approach the same concerns.

The three activation pathways identified in this report each require a different set of boundaries. The following are not exhaustive. They are the conditions most clearly warranting response — cases where the deployment of C2PA infrastructure crosses from legitimate use into surveillance harm, and where the accountability mechanisms described above should be empowered to act.

Legislative and regulatory misuse

Mandating government-issued identifiers — national ID numbers, biometric identifiers, press credential numbers — in Content Credentials as a condition of content creation or distribution is incompatible with press freedom and freedom of expression. Requiring credentialing as a condition of platform distribution in contexts where credentialing is not genuinely voluntary for all creators converts an authenticity tool into a licensing mechanism. Both conditions should trigger mandatory assessment and public findings.

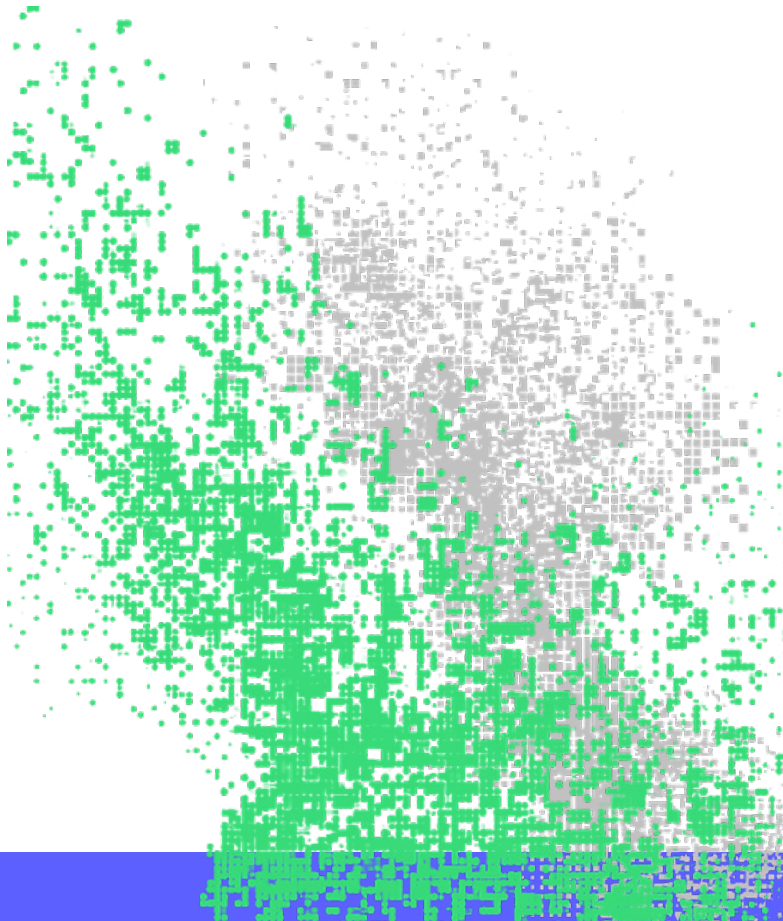
Architectural capture

Where a state requires creators to obtain identity credentials from a state-controlled authority as a condition of producing or distributing credentialed content, two compounding harms follow. That authority becomes a gatekeeper: only creators it approves can produce content that platforms recognize as credentialed, effectively converting an authenticity infrastructure into a licensing regime without requiring any licensing legislation. It simultaneously becomes a surveillance registry: the authority holds, without any requirement for judicial process, a direct record linking every credentialed creator to every piece of content they have produced.

The C2PA's accountability tools do not reach that authority directly. What they do reach are the tools that creators use to sign their content — and that accept those state-issued credentials as a condition of doing so. Tools operating in deployment contexts where state-controlled identity credentials are required should trigger mandatory assessment. Where a tool is itself operating under a legal mandate it cannot refuse, removing it from the conforming products list is an incomplete remedy. Public findings and direct advocacy against the mandate are necessary complements. The most egregious form of architectural capture is also the one existing accountability tools are least equipped to address alone — and that gap is itself worth naming publicly.

Design and implementation failure

Not all harm requires a hostile actor. Tools that embed identity disclosure at the point of content creation without meaningful opportunity for creators to understand or limit that disclosure, and validators that generate identifiable records of viewer behavior without disclosure or any means of refusal, represent failures of design that the accountability mechanisms described above should be equipped to assess. Negligence and intent are different problems, but their consequences for the people this report describes can be identical.



Closing Note and Call to Action

A Note for Regulators

Content provenance is a key mechanism for restoring trust in the information environment. Getting the regulatory design right — rather than simply mandating it — is what determines whether it serves that purpose. Regulators who engage seriously with the technical architecture, who consult with civil society and affected communities before frameworks mature, and who build in review mechanisms as the technology evolves, are more likely to develop solid foundations. The risks documented in this report are an argument for regulating thoughtfully, with the full picture in view. They are also an argument for looking beyond the C2PA: the surveillance risks documented here attach to content provenance and authenticity infrastructure broadly, and regulatory frameworks that reference specific standards without addressing those risks at the governance layer will encounter the same problems regardless of which standard they mandate.

A Note for the Public Interest Community

The infrastructure being built to verify digital content will shape what can be published, trusted, and contested. It is being built now largely without civil society input. The implications for journalists, human rights defenders, and communities documenting abuse are significant and underappreciated. Engaging in the C2PA's ecosystem and in legislative processes referencing provenance standards is essential. More specifically: the C2PA has both the opportunity and the standing to take the lead in addressing these risks within its own ecosystem, and civil society has a role in making that case directly to the C2PA community. Calling for independent oversight mechanisms and accountability structures — within the C2PA, across the content provenance field, and over emerging content transparency systems more broadly — before deployment norms are locked in, is where engagement is most urgently needed.

***

Designing and governing the C2PA well is not only a matter of protecting the information ecosystem — it is an opportunity to demonstrate that a technical standards community can take surveillance risks seriously and build accountability structures that match the scale of the infrastructure it is creating. Content provenance can only work if people trust it. That trust depends on the infrastructure being used for what it is designed for, and on there being credible, visible mechanisms for identifying and responding when it is not. The C2PA is best positioned to help build those mechanisms. Surveillance misuse that goes unnamed, unaddressed, or normalized will erode that trust in ways that no technical improvement can recover. The value of the ecosystem depends on it being governed as seriously as it is built.

References

[1] Coalition for Content Provenance and Authenticity (C2PA). C2PA Technical Specification, version 2.4. C2PA Specifications. Available at: https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html

[2] Coalition for Content Provenance and Authenticity (C2PA). C2PA: Providing Origins of Media Content. c2pa.org. Available at: https://c2pa.org/

[3] Coalition for Content Provenance and Authenticity (C2PA). C2PA Conformance Program. Available at: https://c2pa.org/conformance/

[4] Coalition for Content Provenance and Authenticity (C2PA). C2PA Conformance Explorer. Available at: https://spec.c2pa.org/conformance-explorer/

[5] Coalition for Content Provenance and Authenticity (C2PA). C2PA Membership. Available at: https://c2pa.org/membership/

[6] Coalition for Content Provenance and Authenticity (C2PA). C2PA Harms Modelling, version 2.4. C2PA Specifications. Available at: https://spec.c2pa.org/specifications/specifications/2.4/security/Harms_Modelling.html

[7] Coalition for Content Provenance and Authenticity (C2PA). User Experience Guidance for Implementers, version 2.2. C2PA Specifications. Available at: https://spec.c2pa.org/specifications/specifications/2.2/ux/UX_Recommendations.html

[8] Coalition for Content Provenance and Authenticity (C2PA). Guidance for AI and Machine Learning, version 2.2. C2PA Specifications. Available at: https://spec.c2pa.org/specifications/specifications/2.2/ai-ml/ai_ml.html

[9] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Article 50. Official Journal of the European Union, L 2024/1689. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689

[10] European Commission. Code of Practice on Marking and Labelling of AI-Generated Content, Second Draft. EU AI Office, 5 March 2026. Available at: https://digital-strategy.ec.europa.eu/en/library/commission-publishes-second-draft-code-practice-marking-and-labelling-ai-generated-content

[11] California Legislature. Senate Bill 942: California AI Transparency Act. Approved by Governor 19 September 2024. Available at: https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB942

[12] California Legislature. Assembly Bill 853: California AI Transparency Act (amendments). Approved by Governor 13 October 2025. Available at: https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260AB853

[13] California Legislature. Assembly Bill 2713. 2025–2026 Regular Session. Available at: https://legiscan.com/CA/text/AB2713/id/3423548

[14] California Legislature. Senate Bill 1000. 2025–2026 Regular Session. Available at: https://legiscan.com/CA/text/SB1000/id/3353839

[15] Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, and State Administration of Radio and Television. Measures for Labeling of AI-Generated Synthetic Content (State Information Office Tongzi [2025] No. 2). Promulgated 7 March 2025, in force 1 September 2025. English translation available at: https://www.chinalawtranslate.com/en/ai-labeling/

[16] Ministry of Electronics and Information Technology, Government of India. Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules 2026, G.S.R. 120(E). In force 20 February 2026. Available at: https://egazette.gov.in/WriteReadData/2026/269993.pdf

[17] Alimardani, M., Castellanos, J., and Martins dos Santos, B. (2026). India Bets on AI Detection. Every Regulator Should Watch What Happens Next. Tech Policy Press, 18 February 2026. Available at: https://techpolicy.press/india-bets-on-ai-detection-every-regulator-should-watch-what-happens-next/

[18] Creator Assertions Working Group (CAWG). CAWG Identity Assertion. Available at: https://creator-assertions.github.io/identity/

[19] Coalition for Content Provenance and Authenticity (C2PA). Soft Binding API, version 2.2. C2PA Specifications. Available at: https://spec.c2pa.org/specifications/specifications/2.2/softbinding/Decoupled.html

[20] Parsons, A. (2024). Durable Content Credentials. Content Authenticity Initiative, 8 April 2024. Available at: https://contentauthenticity.org/blog/durable-content-credentials

Published by WITNESS

witness.org

Author: Jacobo Castellanos Rivadeneira

Contact: jacobo@witness.org

This report is a companion to the research paper C2PA Content Credentials and the Surveillance Risk: Adversarial Scenarios and Governance Gaps in the Content Provenance Ecosystem by the same author.

Suggested citation: Castellanos, J. (2026). C2PA Content Credentials and the Surveillance Risk: Adversarial Scenarios and Governance Gaps in the Content Provenance Ecosystem. WITNESS.

This report is published by WITNESS, a civil society organization that has used video and technology to support human rights defenders since 1992. WITNESS is a member of the Coalition for Content Provenance and Authenticity (C2PA) and co-chairs its Threats and Harms Task Force, the body responsible for the C2PA's own harm assessment work. That position gives WITNESS direct knowledge of the specifications, the governance structures, and the communities the standard is meant to serve — and creates an obligation to speak clearly when the infrastructure carries risks those communities have not been adequately warned about. This report is an exercise of that obligation. It reflects the independent analysis of its author and does not represent the position of the C2PA or its membership.

This report was written with support from the Digital Democracy Initiative.

Digital Democracy Initiative

WITNESS