Content Credentials: Trust Issues, Technical Solutions and Future Perspectives Using Encrypted Metadata in Image Processing

György Wersényi* and Victor Koech*

***Abstract*—Emerging technologies offer validation and authentication solutions in the field of audiovisual content creation. Visible or invisible watermarking, embedded metadata, and digital signatures can be used to maintain the validity and creditability of still images and video data. The Coalition for Content Provenance and Authenticity (C2PA) was established to create an open source framework and to provide technical solutions for image capture, processing, delivery, and verification. The leading market players in hardware and software development set the goal of applying encrypted metadata information to guarantee the authenticity of the data. Currently, only a few devices and applications are available and have been implemented based on this technology. This paper gives an introductory overview of the recent state, highlighting advantages, drawbacks, available implementations, and future perspectives on research directions.**

***Index Terms*—Content Credentials; Validation; Deepfake; Image Processing**

I. INTRODUCTION

In the age of rapid propagation of digital media, the truthfulness of the media has become increasingly important. Social networks have devolved journalistic responsibilities to the masses by making it possible for anyone with any motive to have unregulated access to audiences far and wide [1], [2]. This concern is also enhanced by the increased ease of access to photo-manipulation tools, some of which even include automation to lower the previously high technical bar [3]. Additionally, we have now entered the era of generative artificial intelligence (AI), which democratizes the ability to create synthetic content that has a convincing resemblance to real media. These developments have propagated and increased the complexity of disinformation and misinformation, which has caused severe challenges in determining the origin and authenticity of content [4], [5]. Malevolent people can take advantage of these tools to mislead, impersonate, and even conduct cyber-crimes which, in turn, erodes public confidence in systems such as governments, broadcast media and even personal communication [6]. This risk has resulted in an environment that requires verification of the origin and authenticity of any given digital content. A recent study on AI in the music industry revealed that 97% of people cannot tell the difference between fully AI-generated and human made music [7]. Furthermore, there was an overwhelming support for labeling of 100% AI-generated music, and more than half of the subjects felt uncomfortable by not being able to tell the difference.

* Széchenyi István University, Győr, Hungary (e-mail: wersenyi@sze.hu, koech.victor@sze.hu)

A. Trust and Validation

The rise of fake and AI-generated images poses major challenges to trust and validation in media, politics, and social networks [8]. Deepfakes and synthetic visuals can spread misinformation, manipulate public opinion, and undermine trust in authentic content. Such images are prevalent in journalism, advertising, and social media [9]. Technologies such as watermarking, blockchain-based provenance tracking, and encrypted metadata can help authenticate sources and restore trust in visual information [10], [11].

Key technological solutions to improve image authenticity and validation focus on provenance tracking, digital watermarking, cryptographic hashing, and AI-based forensic detection [12]–[16].

- Digital watermarking embeds invisible identifiers into images, allowing origin verification and tamper detection;
- Blockchain-based provenance systems record immutable metadata about image creation and modifications;
- Cryptographic signatures and hashing verify file integrity by comparing digital fingerprints; and
- AI forensics detect manipulations using pixel-level inconsistencies or generative adversarial network fingerprints.

B. Definition of Content Provenance

The foundation of the content provenance model is built on the understanding that there is a constant whack-a-mole going on between generative AI models and AI detection models, which is not sustainable. As generative AI models continue to improve, the detection of their outputs also becomes more complex, and the detection models have to play catch up [17].

DOI: 10.36244/ICJ.2026.1.7

The content provenance approach therefore tries to navigate around this problem by creating a parallel ecosystem of trust which shifts the burden of proof from the content consumer to the content producer. Content creators and publishers are the ones who are empowered to offer proof of their content history, rather than regular consumers who are skeptical and concerned about the trustworthiness of the content they encounter [5].

C. The Coalition for Content Provenance and Authenticity

According to the Coalition for Content Provenance and Authenticity explainer, C2PA is a conglomerate of major technology companies, media entities, and stakeholders spanning multiple industries with the sole objective of promoting an open and universal technical standard to build digital trust [5]. The coalition's main product is the C2PA technical specification, which outlines a framework for generating and embedding content credentials into digital media. The credentials serve as a secure electronic metadata package that is tied to the digital media. The CSI-Content Credentials report compares these credentials with nutrition labels that provide the consumer with details of the ingredients used and the origins of the products [18]. Inspection of these credentials is therefore meant to satisfy the curiosity of where the content originated, what tools were involved, and what manipulations were made over its entire history.

This paper provides a review of the C2PA framework and its applications. It presents the technical architecture, the evolution of its ecosystem, its security stance compared to the available technical specifications, the latest industry developments, and future perspectives in research. The evaluation will also highlight the limitations of the framework, the relationship with other alternative authentication technologies, and its implications on society and ethics. The paper is structured as follows. Section 2 introduces the C2PA framework, followed by basic technical specifications in Section 3. Section 4 deals with adoption and impact; Section 5 discusses limitations. Finally, privacy and ethical issues, as well as a comparison and outlook in development and research will be highlighted.

II. The Emergence and Evolution of C2PA

The C2PA coalition was officially established in the year 2021 as a Joint Development Foundation housed under the Linux Foundation. This happened as a result of the consolidation of two parallel initiatives, which had also been keen on solving the problem of digital provenance. This merging was a welcome change in the re-evaluation of a fragmented standards landscape, while it also helped bring together the expertise and resources of key industry players [19]. The parallel initiatives were The Content Authenticity Initiative (CAI) founded in 2019 by Adobe and Project Origin, also founded in 2019, by Microsoft and BBC. The former was mostly driven by creators with a desire to claim authorship of their work, while the latter was driven by the broadcast industry to counter disinformation and misinformation in the news ecosystem.

A. Governance of C2PA

The C2PA coalition is led by a committee of its bigger members, including Adobe and Microsoft who oversee the software, Arm and Intel who oversee the hardware and chip design, BBC who represent the media, and Truepic who specialize in the authentication technology. This group ensured that there would be good consideration for the entire life cycle of digital media from the point of capture to the editing and finally to its distribution [20]. On a more positive note, there has been a significant expansion of the steering committee, with more industries getting representation. Some of the newer members include early adopters like Sony and X, advertisers such as Publicis Groupe, and AI technology companies such as Google, Amazon, Meta, and OpenAI [21]. The involvement of big players from various industries highlights the coalition's deliberate strategy to expand the scope of the proposed solution. The problem of misinformation and disinformation is large and a solution by one entity would not be able to cover the vast landscape of the Internet [19]. This grouping therefore goes some way to bring about a network effect which can help bring about ubiquitous adoption which is open and interoperable.

B. The C2PA Charter

All work carried out by the C2PA working groups is governed by its official charter, which is a set of core principles framed as the constitution of the specification. The principles including interoperability, privacy, simplicity, global applicability, performance, prevention of abuse, and unbiased viewpoint are meant to ensure that the development is technically sound and ethical [5]. The principles are also designed to help with the decision-making of the working groups, so that all the builds undergo a set of checks through security reviews and harms modeling exercises before they are publicly implemented [22].

III. Technical Specifications and Architecture

A. Overview and Core Aspects of C2PA

To ensure that a digital asset's history is secure, tamper-evident, and verifiable, the C2PA framework layers on top of an architecture of data structures and cryptographic processes. According to the latest C2PA specification, version 2.2, the data model is based on hierarchy, by which the provenance information is segmented into specific components as follows [23]–[25].

1) Assertions: These are fundamentally the identifying units within the C2PA framework as they define a fact about a digital asset. They are all labeled using namespaced strings, such as c2pa.actions, which describe the actions that have been performed on the asset, or c2pa.ingredient that is linked to other assets that were used in the formation of the digital chain. These assertions are defined by C2PA or other entities.
2) Claim: This data structure ties all the assertions together to a single event in time. The claim is digitally signed

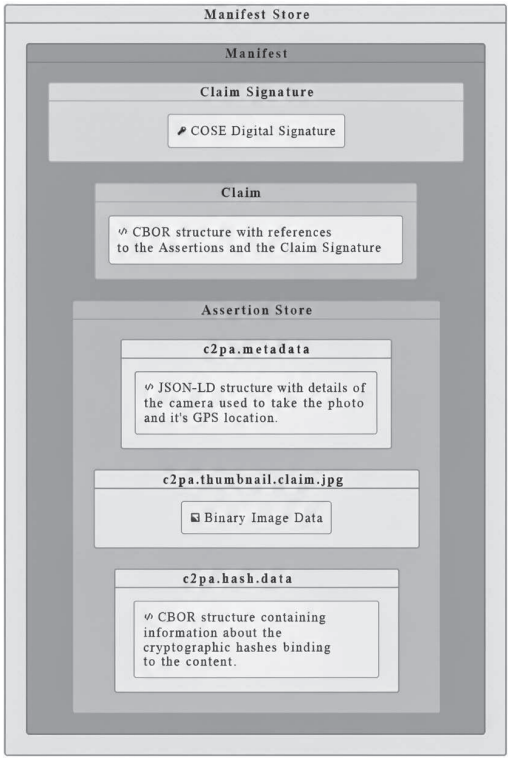

Fig. 1. C2PA Manifest of an image [24].

by the producer/modifier of the asset, cryptographically hard binding all the assertions to indicate their validity.

3) Manifest: This is the container of the verifiable asset provenance. It contains a single claim, the signature of the claim, and a set of assertions tied to the claim. This data package closely resembles the nutrition label of the asset. There can be multiple manifests in each asset that represent each tool that contributed to its lifecycle. Figure 1 shows a manifest of an image.
4) Manifest Store: This overarching container holds all the manifests associated with an asset. The store is a representation of the complete and cumulative provenance history of an asset and can be embedded in the asset file or hosted externally and linked to the asset.

B. Authenticity and Integrity

Establishing the trustworthiness of C2PA content credentials is based on cryptographic mechanisms that ensure that data can be verified to be authentic and have not been tampered with. These mechanisms are as follows.

1) Hashing: This is a hard-binding technique that creates a tamper-evident trace between a digital asset and its manifest. It involves generating a cryptographic hash that is tied to the bytes of the digital asset and is stored in an assertion contained in the manifest [25]. The hash is generated using any of the standard algorithms such as SHA-384 or SHA-256 and this generally results in a form of digital fingerprint of the asset which definitively determines the tamper status of the asset during validation [26].
2) Signing: The manifest containing the hash also contains a claim that is digitally signed by an actor that can either be a software application or a hardware device that creates it. This signing uses a public key infrastructure model in which the signature to the claim uses a private key that can be universally verified using the corresponding public key. Public keys are hosted by certificate authorities on their digital certificates (typically X.509 certificates) and show the creator and a tamper status since their creation [27].
3) Timestamping: For good record-keeping and chaining, details of when a manifest was signed are captured in the form of a time stamp. The signing process utilizes the services of a time stamp authority to tie the hash of the claim with a dated token. The token is bundled with the manifest and provides proof that the signature was created on the said date and time [26]. This can then be used to later verify the validity of the certificate according to the expiration or revocation status.

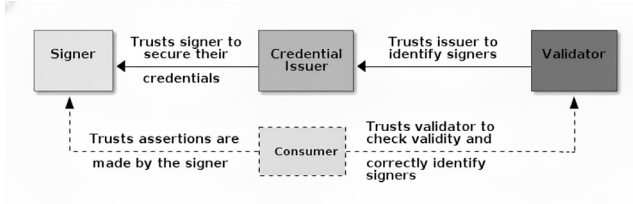

Fig. 2. The C2PA Trust Model [24].

C. The C2PA Trust Model

The framework aims to provide verifiable information about an asset's provenance so that the basis of trust is anchored in the identities of the signers. Devices and software that perform content credential validation do this by checking the integrity of the provenance data attached to the digital asset, so this is not just a check for 'realness' or 'fakeness'. There is a Trust List made up of certificate authorities whose certificates are known and acknowledged by the validating application or device. Therefore, the validation process tries to establish whether the certificate used originated from a certificate authority that is on the trust list. Depending on the results of the integrity checks, the manifest can be assigned a state such as valid, trusted, well-formed, or unknown [28]. To achieve the highest level of technical trust, an asset needs to have a well-designed manifest that meets the C2PA specification. The digital signature of the manifest should also be successfully validated as tamper-free, therefore valid, and finally the signature certificate for the manifest chain should be present on the trust list of the validator. Figure 2 shows the so-called trust model within C2PA.

Another aspect of the trust model is the binding of a manifest store with its associated asset either by embedding

the store in the asset file structure or storing the manifests externally in the cloud or in a distributed ledger [26]. Typically, in the embedding approach, JPEG Universal Metadata Box Format containers were used, ensuring that the provenance data moves with the asset as a self-contained file. In some other cases, such as when uploading files to social media apps whereby metadata can be stripped, C2PA advocates for using the external repository method where a soft binding is applied to relink assets to their stripped metadata. The soft binding identifier can be a digital watermark or a perceptual hash that visually identifies the asset or similar content [25].

C2PA uses cryptographic hashing and digital signatures to bind provenance to content in a tamper-evident way. Hashes (i.e., SHA-256) compute content digests over defined byte ranges; signatures over CBOR-serialized claims use X.509-based credentials via COSE (e.g., ECDSA, EdDSA, RSASSA-PSS). In the trust model, decisions rest on the identity of the signer (credential chain anchored in trusted roots) and validity of their certificate. Validators maintain trust lists after approval, check revocation, and assess whether a signing credential was valid at the time of signing. Identity assertions may include devices, applications, or pseudonyms. Information security and threat modeling outline ongoing threat modeling, encourage design against abuse, and emphasize key management, minimizing key reuse, revocation strategy, and securing claim generation systems. It also highlights the evaluation of harms, misuse and abuse as a continuous process that respects privacy, human rights, and evolving threats.

D. Privacy and Ethics

The foundation of C2PA is to ensure that transparency is brought to the forefront; however, this conflicts with the fundamental right to privacy. If the history of a digital asset is to be detailed and unalterable for accountability purposes, everything including who created it, where, when, and how should be included, but this will also bring about concerns of surveillance, misuse, and speech suppression [28]. These are concerns that the C2PA foundation is aware of and has taken some steps toward addressing. In the Harms Modeling documentation, C2PA outlines the possible harms that could occur, and the mechanisms by which creators can take control.

One of the key tools by which the C2PA tries to address concerns is by preaching that the whole project is voluntary and that no content should have credentials attached by default. This means that one must consent to the recording of provenance data, even if they are using a C2PA enabled tool [5]. To complement this, creators are also provided with a redaction feature which allows them or subsequent editors to subtract some assertions such as those containing sensitive information without invalidating the attached manifest. This subtraction leaves a record to ensure that transparency is preserved, but sensitive information is permanently removed [29]. Lastly, it is not a requirement for the manifest signature to identify the signer because pseudonyms or anonymous certificates can be used to protect vulnerable people like whistle-blowers and activists [30].

The Harms Modeling framework has also formalized a governance process in which potential social issues and their impacts are accounted for. The idea is to anticipate how technology might be used to negatively affect stakeholders and the world at large, by classifying intended users and their use cases, weighing the potential harms and creating countermeasures such as changes to specifications, design recommendations, and public interest campaigns [31]. These safeguards do not necessarily address all issues because as the adoption of C2PA increases, there is an ethical challenge that arises, whereby content lacking the defined identifier could be deemed suspicious in the future and this could penalize creators who have legitimate reasons not to adopt it [28].

E. C2PA versus Competing Technologies

There are several technologies that have been specifically built to address the issue of digital trust, but each of them has their shortcomings. When comparing C2PA with AI-based deepfake detection solutions, the most notable distinction is in the approach, which is proactive in the former and reactive in the latter. C2PA's approach is to establish the content's authenticity from the point of creation, which leaps ahead of the inherently delayed approach of trying to verify manipulation through AI [27], [32]. Although fake AI detection techniques might seem limited, especially considering their propensity to perform well with known manipulations but fail on new ones, it still serves as the best approach for the analysis of existing content created before C2PA [33].

When comparing C2PA with digital watermarking technologies, there is more of a synergetic relationship, whereby watermarks complement metadata packages. C2PA manifests can be easily removed from the digital asset in transmission while the digital watermarks remain attached to the pixels of the asset and therefore can survive a variety of transformations, including screen captures [34]. This has been the foundation of the Durable Content Credentials model, which combines the full external CPA manifest with a robust digital watermark bundled within the content itself. The resulting digital asset can be identified by a C2PA-compliant application even if the metadata had been removed, which means that the asset can have its provenance information restored [35].

C2PA also has a complementary relationship with perceptual hashing through the use of cryptographic functions to establish bindings. Hashes such as SHA-256 are known to be very profound, such that any change to the file, no matter how small, will result in a change in the hash value, which makes them ideal for bit-for-bit integrity but powerless for similarity comparisons. With perceptual hashes, the hash produced can be similar for content that is visually comparable despite changes to file size or minor alterations. They are therefore suitable for identifying copies or close duplicates of a digital asset shared on the Web [36]. C2PA has used perceptual hashing technology to establish soft bindings so that when manifests are detached, the hash can be used to find original or similar assets and reconnect them with provenance information [37].

Blockchain systems are also closely related to C2PA in that they are used to create immutable records, albeit with a different underlying architecture. For C2PA, cryptographic metadata can be embedded directly into the file for offline use, eliminating the need for a connection to a distributed network [38]. This means that the specification can utilize blockchain, but it is not necessary [26]. The advantage of using the decentralized distributed ledger of blockchain technology is that reliance on a single authority is removed, leading to extreme resistance to manipulation. Projects like the Numbers Protocol have combined these two technologies such that the C2PA provenance data are signed and the hash of the manifest is registered on a blockchain [39].

IV. DISCUSSION

Ensuring authenticity and trust in images and videos is critical to preserving the integrity of digital information. Manipulated or synthetic media can distort truth, fuel misinformation, and erode public confidence. Reliable provenance, cryptographic validation, and transparent metadata are essential to maintain accountability and verifiable trust in the dissemination of visual content. However, creating multi-metric evaluations combining robustness (survivability), security (tamper resistance, key compromise scenarios), utility (verification latency), and human-centric metrics (trust, comprehension) is a difficult task.

Providing open datasets and tools (scripts that apply common transforms, edits, platform-like re-encodings) are needed for others to reproduce results. Adversarial testbeds can simulate common distribution chains from the camera through the editor and the distribution network to the final result (screenshot/re-upload). Threat models must be defined explicitly to show what attackers can and cannot do (e.g., full access to platform storage vs. passive network attacker).

A C2PA manifest acts as a verifiable container of provenance data, enabling users and systems to confirm the authenticity and history of the modification of digital content. Furthermore, a C2PA claim is a trusted signed declaration embedded in the manifest that documents verifiable facts about the lifecycle or attributes of digital content. Schemas make provenance data understandable and consistent, while signatures make them trustworthy and tamper-proof. A schema is a descriptive data model that defines what data looks like (structure and validation); a signature is a cryptographic mechanism that verifies who created or altered data (authenticity and integrity). Together, they target a transparent, interoperable, and secure ecosystem.

A. Hardware Support

Currently, only a few vendors offer models with C2PA authentication (see Figure 3). These are available either out-of-the-box in hardware or via firmware/software updates. Some are partial or require licensing, and some vendors have announced intentions or are working on updates, but those features may not yet be available broadly.

Vendor	Camera Models / Details	Notes / Status
Leica	Leica M11-P, M11-D, SL3-S	Built-in C2PA provenance signatures.
Sony	A1 II, A1, A9 III, A7S III, A7 IV	Firmware-based C2PA; may need licensing.
Nikon	Z6 III	C2PA support via firmware rollout.
Fujifilm	X-T50, GFX100S II	Supports C2PA authenticity metadata.
Canon	EOS R1, EOS R5 Mark II	Native or firmware C2PA support.

Fig. 3. Vendors and models.

The movement officially started with Leica, who pioneered the M11-P camera in 2023 that featured built-in support for content credentials, and this was soon followed by another model, the SL3-S, in 2025 with the same capability [20]. This paved the way for other camera manufacturers such as Nikon, Canon, Panasonic, and Sony to join the initiative [27], [40]. Another significant milestone for hardware adoption was reached when Samsung, the largest smartphone manufacturer, announced that the Galaxy S25 lineup of phones would also adopt native C2PA support. This was shortly followed up by the Google Pixel lineup, also releasing with native support.

It is also possible that authenticity services may be available only in certain countries or regions, or only for certain user classes (i.e., press agencies) initially. Furthermore, even if the camera embeds the metadata, a software is needed in the workflow (editing, export, sharing) that preserves this metadata so the provenance chain remains intact. If metadata is stripped (by certain social media platforms or export tools), the C2PA signature might be lost.

The adoption of technology is evolving; many cameras require a specific firmware version, an optional licensed 'authenticity' service, or registration with the vendor cloud to load signing certificates. Furthermore, chip-level/phone implementations (i.e. Google Pixel 10 and Qualcomm platform integrations) are appearing that use content credentials at capture time when the phone OEM and camera stack implements it.

B. Software Support

A concise and up-to-date rundown of the leading software solutions and toolkits that implement content credentials for images can be assembled, whether they are primarily online (cloud/web), offline (desktop/mobile apps/local libraries), or hybrid services.

Primarily offline desktop apps (with optional online features) include Adobe Photoshop and Lightroom. Primarily online (cloud and APIs) solutions include Truepic (verification and authenticity of APIs), platform/Content delivery network integrations (Cloudflare), and platform verification features (YouTube, Google Search) [41]. Some vendors offer device-based hybrid solutions, i.e. Google Pixel or Sony and Leica, which embed manifests in-camera.

Developer toolkits and off-line libraries that are embedded into applications are available as open-source SDKs and reference tools (c2pa-js, Python examples, etc.). The CAI/C2PA community provides open-source SDKs, the Verify inspector tool, c2pa-js, Python/other examples, and reference implementations (GitHub) so developers can create or validate content

credentials in apps and pipelines. These are used to build server-side and client-side verification. Truepic is the best-known vendor, but other emerging platform vendors also plan to run private implementations. On the distribution side of the digital media, C2PA has also announced that some of the largest players have also begun integrating the framework into their platforms. This has been led by Meta (Instagram and Facebook), LinkedIn, and Tiktok [41].

The following services are currently available:

1) Adobe — Content Credentials (Photoshop, Lightroom, Web tools). Offline desktop apps can embed content credentials at export; online web tools for inspection and management (Adobe Content Authenticity web app, Inspect).
2) Truepic verify and authenticity platform. Its cloud and mobile SDKs provide capture and verification services, device attestation, and an authenticity platform for publishers, platforms, and enterprises.
3) Google Pixel and Photos for platform integrations is a hybrid device-level capture tool (Pixel/Android) that can embed content credentials at capture time (on-device) and online platform/display integrations (Google Photos, Search) to surface provenance.
4) Microsoft offers Project Origin and Azure integrations. It is an online platform and research initiative; developer documents show content credentials support for generated media.
5) Platform and Content delivery network adopters, such as Cloudflare, YouTube, and others are beginning to preserve and expose content credentials (example: Cloudflare added a “Preserve Content Credentials” option for hosted images; YouTube has experimented with C2PA-based labels).

C. Vulnerabilities and Adversarial Robustness of C2PA

The effectiveness of C2PA in achieving its objectives is challenged by several factors briefly mentioned. The main vulnerability is the fragility of embedded metadata, which can be easily removed or altered when uploaded to content delivery networks for optimization or privacy reasons [28]. A much simpler form of this vulnerability is the rampant laundering of content through screenshots and screen recording to create a new file for re-uploading [42]. Outside of these fundamental limitations, security researchers have also pointed out some sophisticated exploits such as provenance piggybacking, where an attacker takes an authentic credentialed asset and layers a manipulated or deep-faked element on top of it [43], and unprotected metadata manipulation, where the unprotected elements like EXIF data can be changed without invalidating the C2PA signature [44]. The RAND corporation has also pointed out that the framework’s threat model needs updating because it has been the same since version 1.0 from January 2022 with significant changes in the landscape driven by generative AI [45], [46]. This criticism highlights a significant privacy dilemma: the inherent push for transparency in C2PA conflicts with the need for safety/privacy for vulnerable creators. Another documented shortcoming of C2PA is that adding manifests to digital assets tends to increase file sizes, so assets that undergo multiple edits can become prohibitively large. This can result in bandwidth and storage problems for low-resource situations [47].

D. Outlook of C2PA

As the C2PA framework evolves and graduates from specification to general adoption, it will be critical that it is adopted by the International Organization for Standardization (ISO). Currently, there is some work to get this done in a fast track process published as ’ISO / DIS 22144, Authenticity of Information - Content Credentials’ [35]. If this happens, then it will carry more weight and increase adoption by governments and more corporations. So far, there has been some normal challenge to the technical details of the framework, but they are minor and do not dispute the core architecture [48].

The technical roadmap of the specification shows that it is focused on addressing several key challenges, such as solving metadata removal through vendor-agnostic watermarking, strengthening the threat model to counter AI attacks, addressing the issue of file size through compression or better manifest management, and achieving end-to-end provenance preservation [26], [46], [49]. There has also been a shift in the digital media landscape, with governments worldwide considering or enacting legislation that mandates transparency. In the United States, for example, Utah and California states have passed laws requiring verification of online content, while the United Nations has also made resolutions encouraging interoperability of authentication mechanisms. [50].

E. Research Directions

Due to the novelty of this topic, there is limited scientific research data available. Most of what is available is tech reports, preprints, repository entries, or non-reviewed publications. The most recent directions include cryptographic provenance, technical standard development, authenticity issues, and ethical frameworks [28], [51]–[53].

The most important field is trust enabling and calibration. Early user studies exist, but are limited. The goal is to experimentally measure how provenance displays affect user judgments for images and short videos across demographics and platforms. Controlled user experiments varying label wording, iconography, provenance depth should be designed in which accuracy in detecting manipulated media and changes in credibility/trust ratings can be tested. Different provenance user interfaces and labels can affect user trust, perception, and behavior (what users understand from “captured with camera and C2PA” vs. “AI-generated” labels). This issue is strongly related to trade-offs between provenance transparency, creator privacy, and identity binding. The support of selective disclosure, pseudonymous attestations, and legal/regulatory constraints is of paramount importance.

Emerging critiques warn against over-promising provenance [54]. Some of the vulnerabilities that have been highlighted, such as practical exploits and a weak threat model, show the

need to restrain enthusiasm until more research is conducted and solutions implemented. Personal data and details about the origin of the image (reliable data about the time and location of the capture, etc.), possible tracing of individuals based on metadata highlight privacy and data security issues. Designing cryptographic protocols that enable content to carry provenance claims without exposing sensitive identity fields, that is, zero-knowledge proofs to assert 'created by a verified source' while hiding identity, is essential to build trust.

Scalability and verification at Internet scale deals with the efficient verification across network edge nodes, browsers, social platforms, and truncation/screenshot scenarios (how to preserve provenance when content is transformed for distribution). For screenshots and re-upload, no concrete proposals have been made yet [55]. Studies have been carried out to determine how effective metadata manifests resist deletion and tampering vs. robust watermarks and ML-based verification [37]. Hybrid schemes that combine C2PA manifests with robust invisible watermarks are a major focus. There is a need to design an architecture that uses C2PA metadata and a perceptually robust watermarking scheme to survive common transforms (recompression, cropping, screenshotting) while preserving unforgeability will lead the way to increase trust.

Practical systems and prototype deployments are emerging regarding the embedding of continuous provenance for live streams, low-latency signing, and secure key handling [56], [57]. Building and measuring C2PA-compatible live streaming prototypes is the next step in development. Another focus could be on interoperability, which refers to the standards, protocols, technologies, and mechanisms that allow data to flow between diverse systems with minimal human intervention [58].

V. CONCLUSIONS

In this paper, we provided a brief overview of the current state of content provenance and authenticity regarding visual media, focusing on still images. The C2PA and the CAI association founded by leading market players paved the way for developments and technical solutions by formulation of an open, royalty-free technical standard that serves as a basis for the C2PA member's efforts against disinformation. Only a few experimental results have been available to date. The research directions and open questions were highlighted. Concerns about data privacy can undermine trust. Key and identity management at scale (who issues keys, revocation) is critical for platform trust. Provenance does not automatically stop misinformation; it can help, but a sociotechnical evaluation is necessary to avoid overpromising. The fine balance between the need for transparency and the right to privacy has introduced an ethical dilemma for the framework to figure out a solution.

The C2PA framework has great ambition and is serving a critical role in restoring trustworthiness to the digital ecosystem. Its success faces some hurdles that will depend on the coalition's resolve to improve the standard to meet the arising needs, the industry's appetite for privacy-preserving implementation, and the general acceptance of provenance by the public with its critical nuances. Future directions in research include the evaluation of human factors, but also technical issues such as cryptography, detection of fake images and malevolent intentions, real-time adaptation of the technology into live streams and the implementation on hardware (cameras) and software application level (social media platforms, search engines, generative and detective AI systems).

REFERENCES

[1] J. Golbeck, "Combining provenance with trust in social networks for semantic web content filtering," in *International Provenance and Annotation Workshop*. Springer, 2006, pp. 101–108.

[2] A. Bernstein and A. Gomila, "The truth in social media," *Topoi*, vol. 44, pp. 127–138, 3 2025. DOI: 10.1007/s11245-024-10039-6

[3] J. Jang, "Self presentation using photo-editing apps on social media," Master's Thesis, Seoul National University, Seoul, South Korea, February 2023. [Online]. Available: https://s-space.snu.ac.kr/handle/10371/194001

[4] E. Ferrara, "Genai against humanity: Nefarious applications of generative artificial intelligence and large language models," *Journal of Computational Social Science*, vol. 7, no. 1, pp. 549–569, 2024.

[5] Coalition for Content Provenance and Authenticity, "C2PA Explainer, v2.2," https://spec.c2pa.org/specifications/specifications/2.2/explainer/Explainer.html, 2025.

[6] S. Kreps and D. L. Kriner, "The potential impact of emerging technologies on democratic representation: Evidence from a field experiment," *New Media & Society*, vol. 26, pp. 6918–6937, 12 2024. DOI: 10.1177/14614448231160526

[7] Deezer, "Deezer/Ipsos Survey: 97% of people can't tell the difference between fully AI-generated and human made music," {https://newsroom-deezer.com/2025/11/deezer-ipsos-survey-ai-music/}, 2025.

[8] R. Chesney and D. Citron, "Deepfakes and the new disinformation war: The coming age of post-truth geopolitics," *Foreign Affairs*, vol. 98, no. 1, pp. 147–155, 2019.

[9] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE journal of selected topics in signal processing*, vol. 14, no. 5, pp. 910–932, 2020.

[10] A. F. Qasim, F. Meziane, and R. Aspin, "Digital watermarking: Applicability for developing trust in medical imaging workflows state of the art review," *Computer Science Review*, vol. 27, pp. 45–60, 2018.

[11] A. Alexander, "Truth in the post-deepfake era: Can blockchain and watermarking restore digital trust?" Available at SSRN 5377851, 2025.

[12] M. Barni and F. Bartolini, Watermarking systems engineering: enabling digital assets security and other applications. Crc Press, 2004.

[13] B. Singh and G. Kasana, "A review of digital watermarking techniques: Current trends, challenges and opportunities," in *Web Intelligence*, vol. 22, no. 4. SAGE Publications Sage UK: London, England, 2024, pp. 523–553.

[14] X. Li, L. Wei, L. Wang, Y. Ma, C. Zhang, and M. Sohail, "A blockchain-based privacy-preserving authentication system for ensuring multimedia content integrity," *International journal of intelligent systems*, vol. 37, no. 5, pp. 3050–3071, 2022.

[15] Q.-u.-A. Mastoi, M. F. Memon, S. Jan, A. Jamil, M. Faique, Z. Ali, A. Lakhan, and T. A. Syed, "Enhancing deepfake content detection through blockchain technology." *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 6, 2025.

[16] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do gans leave artificial fingerprints?" in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2019. DOI: 10.1109/MIPR.2019.00103 pp. 506–511. [Online]. Available: https://ieeexplore.ieee.org/document/8698678

[17] K. Cheng, W. Li, N. Zhang, X. Liu, and H. Wu, "Principles and challenges of generative artificial intelligence detection," *British Journal of Anaesthesia*, vol. 133, no. 4, pp. 899–901, 2024.

[18] Canadian Centre for Cyber Security (CCCS) and NSA and UK NCSC and Australia ACSC, "Content Credentials: Strengthening Multimedia Integrity in the Generative AI Era," *Canadian Centre for Cyber Security, Tech. Rep.*, 2025, version 1.0. [Online]. Available: https://media.defense.gov/2025/Jan/29/2003634788/-1/-1/0/CSI-CONTENT-CREDENTIALS.PDF

[19] C2PA Founding, "C2PA Founding Press Release," https://c2pa.org/c2pa-founding-press-release/, 2 2021.

[20] Content Authenticity Initiative, "How it works," https://contentauthenticity.org/how-it-works, 2025.

[21] "C2PA - Announcements," https://c2pa.org/news/, 2021.

[22] Coalition for Content Provenance and Authenticity, "Guiding Principles," https://c2pa.org/principles/, 2021.

[23] CHESA, "Understanding c2pa: Enhancing digital content provenance and authenticity," https://chesa.com/, 2024.

[24] Coalition for Content Provenance and Authenticity (C2PA), "Content Credentials: C2PA Technical Specification, Version 2.1," C2PA, Tech. Rep. 2.1, 2025, accessed: 2025-10-10. [Online]. Available: https://spec.c2pa.org/specifications/specifications/2.1/specs/C2PA_Specification.html

[25] Coalition for Content Provenance and Authenticity, "Content Credentials: C2PA Technical Specification, v2.2," https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html, 2025.

[26] K. Rathi, S. S. Kumar, and A. N. Mandanna, "Insights into Coalition for Content Provenance and Authenticity (C2PA)," https://www.infosys.com/iki/techcompass/content-provenance-authenticity.html, 2 2024.

[27] M. Demey, "C2PA's Fight Against AI-Generated Deception," https://apryse.com/blog/ai-content-cp2a-authenticity, 11 2023.

[28] World Privacy Forum, "Privacy, Identity and Trust in C2PA: A Technical Review and Analysis of the C2PA Digital Media Provenance Framework," https://worldprivacyforum.org/posts/privacy-identity-and-trust-in-c2pa/, 2025.

[29] Coalition for Content Provenance and Authenticity, "C2PA Implementation Guidance, v2.2," https://spec.c2pa.org/specifications/specifications/ 2.2/guidance/Guidance.html, 2025.

[30] "C2PA Security Considerations, v1.0," https://spec.c2pa.org/specifications/specifications/1.0/security/SecurityConsiderations.html, 2022.

[31] "C2PA Harms Modelling," https://spec.c2pa.org/specifications/specifications/1.0/security/attachments/Harms_Modelling.pdf, 2021.

[32] M. Zorz, "The limits of AI-based deepfake detection – Help Net Security," https://www.helpnetsecurity.com/2024/11/22/ben-colman-reality-defender-deepfakes-detection/, 11 2024.

[33] Y. Xu, P. Terhörst, M. Pedersen, and K. Raja, "Analyzing fairness in deepfake detection with massively annotated databases," *IEEE Trans-actions on Technology and Society*, vol. 5, pp. 93–106, 2 2024. DOI: 10.1109/tts.2024.3365421

[34] A. Parsons, "Durable content credentials," https://contentauthenticity.org/blog/durable-content-credentials, 2024.

[35] U.S. Department of Defense, "Content Credentials: Strengthening Multimedia Integrity in the Generative AI Era," https://media.defense.gov/2025/Jan/29/2003634788/-1/-1/0/CSI-CONTENT-CREDENTIALS.PDF, 2025.

[36] L. Struppek, D. Hintersdorf, D. Neider, and K. Kersting, "Learning to break deep perceptual hashing: The use case neuralhash," in *ACM International Conference Proceeding Series. Association for Computing Machinery*, 6 2022., pp. 58–69. DOI: 10.1145/3531146.3533073

[37] J. Fairoze, G. Ortiz-Jimenez, M. Vecerik, S. Jha, and S. Gowal, "On the difficulty of constructing a robust and publicly-detectable watermark," *arXiv preprint arXiv:2502.04901*, 4 2025.

[38] Content Authenticity Initiative, "Open-source Tools for Content Authenticity and Provenance," https://opensource.contentauthenticity.org/docs/faqs/, 2025.

[39] Documentation for Numbers Protocol Team, "What is C2PA and why do we need it?" https://docs.numbersprotocol.io/introduction, 2025.

[40] J. Tse, "5,000 members: building momentum for a more trustworthy digital world," https://contentauthenticity.org/blog/5000-members-building-momentum-for-a-more-trustworthy_digital-world, 8 2025.

[41] Coalition for Content Provenance and Authenticity, "C2PA News,"https: //c2pa.org/news/, 2024.

[42] Online Brand Ambassadors, "C2pa: Certifying digital media's authenticity," https://www.onlinebrandambassadors.com/c2pa-certifying-digital-medias-authenticity/, 2024.

[43] SCW, "How c2pa can safeguard the truth from digital manipulation," https://www.scworld.com/perspective/how-c2pa-can-safeguard-the-truth-from-digital-manipulation, 2025.

[44] The Hacker Factor Blog, "C2pa from the attacker's perspective," https://www.hackerfactor.com/blog/index.php?/archives/1031-C2PA-from-the-Attackers-Perspective.html, 2024.

[45] RAND Corporation, "Overpromising on Digital Provenance and Security," https://www.rand.org/pubs/commentary/2025/06/overpromising-on-digital-provenance-and-security.html, 2025.

[46] A. Locker, C. Heitzenrater, and T. Helmus, "Overpromising on digital provenance and security," https://www.rand.org/pubs/commentary/2025/06/overpromising-on-digital-provenance-and-security.html, 6 2025.

[47] Infosys, "Insights into coalition for content provenance and authenticity (c2pa)," https://www.infosys.com/iki/techcompass/content-provenance-authenticity.html, 2024.

[48] "ISO/DIS 22144, Authenticity of Information - Content Credentials," 2025.

[49] G. Huszar, "Why Broadcasters Must Embrace C2PA for Content Trust," https://onediversified.com/insights/blog/c2pa, 2025.

[50] Coalition for Content Provenance and Authenticity, "C2PA NIST Response," https://downloads.regulations.gov/NIST-2024-0001-0030/attachment_1.pdf, 2024.

[51] P. Laskar, "Cryptographic Provenance and the Future of Media Authenticity: Technical Standards and Ethical Frameworks for Generative Content," *Journal of Computer Science and Technology Studies*, vol. 7, no. 6, pp. 967–972, 2025.

[52] E. Bureacă and I. Aciobănit, ei, "A Blockchain Blockchain-based Frame-work for Content Provenance and Authenticity," in *2024 16th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. IEEE, 2024, pp. 1–5.

[53] C. Trattner, S. L. Forstner, A. Starke, and E. Knudsen, "C2PA Provenance Labels Increase Trust in News Platforms Across Western Countries," https://www.researchgate.net/publication/391419247 C2PA Provenance Labels Increase Trust in News Platforms Across Western Countries, 5 2025.

[54] A. R. Locker, C. Heitzenrater, and T. C. Helmus, "Overpromising on Digital Provenance and Security," https://www.rand.org/pubs/commentary/2025/06/overpromising-on-digital-provenance-and-security.html, RAND School of Public Policy, Tech. Rep. Note, 2025, accessed: 2025-10-10.

[55] J. Seynhaeve, "An introduction to C2PA — Certifying digital media's authenticity," https://www.onlinebrandambassadors.com/c2pa-certifying-digital-medias-authenticity/, 11 2024.

[56] M. Mesa-Simón, A. Escobar-Molero, B. Sáez-Mingorance, D. P. Morales, J. A. Álvarez-Bermejo, and F. J. Romero, "Enabling live video provenance and authenticity: A c2pa-based system with tpm-based security for livestreaming platforms," *Authorea Preprints*, 2025.

[57] S. Petrangeli, H. Wang, M. Fisher, D. Kozma, M. Mahamli, P. Blumenthal, and A. Parsons, "Integrating content authenticity with dash video streaming," in *Proceedings of the 15th ACM Multimedia Systems Conference*, 2024, pp. 492–498.

[58] J. C. Simmons and J. M. Winograd, "Interoperable provenance authentication of broadcast media using open standards-based metadata, watermarking and cryptography," *arXiv preprint arXiv:2405.12336*, 2024.

György Wersényi was born in 1975 in Győr, Hungary. He received his MSc degree in electrical engineering from the Technical University of Budapest in 1998 and PhD degree from the Brandenburg Technical University in Cottbus, Germany. Since 2002 he has been member of the Department of Telecommunications at the Széchenyi István University in Győr. From 2020 to 2022 he was the dean of Faculty of Mechanical Engineering, Informatics and Electrical Engineering, as well as the scientific president of the Digital Development Center at the university. Currently, he is a full professor, member of the European Acoustics Association (EAA) and the Audio Engineering Society (AES). His research focus is on acoustic measurements, virtual and augmented reality solutions, sonification, cognitive infocommunications, and assistive technologies.

Victor Koech was born in 1991 in Kenya. He received his MSc degree in 2020 in Computer Information Systems at the School of Science and Technology of Kenya Methodist University. He has been a postgraduate PhD candidate since 2025 at Széchenyi István University, Hungary.