Origin Lens: Reclaiming Trust on the AI-Mediated Web Through On-Device Image Provenance Verification*

Alexander Loth
Frankfurt University of Applied Sciences
Frankfurt am Main, Germany
Doctoral Center for Applied Computer Science (PZAI)
Germany
alexander.loth@stud.fra-uas.de

Dominique Conceicao Rosario
Peter Ebinger
Martin Kappes
Frankfurt University of Applied Sciences
Frankfurt am Main, Germany
dominique.conceicaorosario@stud.fra-uas.de,peter.ebinger@fra-uas.de,kappes@fra-uas.de

Marc-Oliver Pahl
IMT Atlantique, UMR IRISA, Chaire Cyber CNI
Rennes, France
marc-oliver.pahl@imt-atlantique.fr

Abstract

Generative AI enables synthetic visual media at a scale that exceeds human verification capacity. This widens an asymmetry on the Web: users encounter images of uncertain provenance with few practical means of assessment. We present **Origin Lens**, a privacy-first mobile framework that performs cryptographic image provenance verification and AI detection entirely on-device. Built on a Rust/Flutter hybrid architecture, the system combines multiple verification signals—C2PA content credentials, generative model fingerprints, and optional retrieval-augmented context—to provide users with graded confidence indicators at the point of consumption. Unlike centralized platform moderation, this approach allows individuals to verify visual content independently. We discuss how client-side verification infrastructure relates to EU regulatory frameworks (AI Act, DSA) and can complement platform governance for information integrity on the Web.

CCS Concepts

• **Security and privacy** → **Digital rights management**; • **Computing methodologies** → *Computer vision*; • **Information systems** → *Multimedia information systems.*

Keywords

Web Trust, Information Integrity, Image Provenance, C2PA, Visual Misinformation, Privacy-First, On-Device Verification, Generative AI, Web Governance, Socio-Technical Systems

ACM Reference Format:
Alexander Loth, Dominique Conceicao Rosario, Peter Ebinger, Martin Kappes, and Marc-Oliver Pahl. 2026. Origin Lens: Reclaiming Trust on the AI-Mediated Web Through On-Device Image Provenance Verification. In *18th ACM Web Science Conference (WebSci Companion '26), May 26–29, 2026, Braunschweig, Germany*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3795513.3806658

*Accepted at the 18th ACM Web Science Conference (WebSci Companion '26). DOI: 10.1145/3795513.3806658

WebSci Companion '26, Braunschweig, Germany
© 2026 Copyright held by the owner/author(s).
This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *18th ACM Web Science Conference (WebSci Companion '26), May 26–29, 2026, Braunschweig, Germany*, https://doi.org/10.1145/3795513.3806658.

1 Introduction

The Web has become a major infrastructure through which people encounter news and participate in public discourse. Generative AI introduces a *verification asymmetry* into this ecosystem: synthetic visual media can be produced at scale, yet individual users lack practical tools to assess content authenticity at the point of consumption. This asymmetry undermines trust on the Web—a recurring concern in Web Science research [35]. Prior work in Web Science has examined how the Web's architecture shapes information flows and vice versa, motivating user-facing verification tools. Large language models have dual-use potential—enabling new disinformation capabilities while offering detection solutions [21]—yet existing mitigation approaches predominantly rely on centralized, platform-level interventions [13]. This creates dependencies on opaque moderation systems and raises surveillance concerns [1], while human factors research indicates that users benefit from immediate, contextual verification signals rather than delayed fact-checks [41].

This paper presents **Origin Lens**, an open-source, privacy-first mobile framework that moves image verification from centralized platforms to individual users.[1] Implementing the C2PA (Coalition for Content Provenance and Authenticity) standard [3] alongside heuristic AI detection, the framework makes provenance verification accessible to non-expert Web users. The contribution is threefold: (1) a privacy-preserving architecture performing verification entirely on-device, (2) a defense-in-depth pipeline combining cryptographic, heuristic, and contextual signals with graded confidence indicators, and (3) a discussion of how client-side verification tools can complement platform governance and regulatory frameworks.

2 Related Work

Recent work in information integrity has focused on robustness evaluation and scalable detection. The *OpenFake* dataset [18] provides deepfake benchmarks, while *Veracity* [6] demonstrates retrieval-augmented fact-checking. Thibault et al. address detection under distribution shift [39]. Technical detection must be paired with effective human-AI interaction. Lin et al. show psychological inoculation has limited real-time effectiveness [41]. Prior JudgeGPT/RogueGPT studies [22, 25] reveal a perception-accuracy gap [24], and empirical

[1] Source code: https://github.com/aloth/origin-lens. Available on the iOS App Store: https://apps.apple.com/app/id6756628121. Android support is planned.

analysis shows cognitive fatigue degrades fake content detection by 10.2 percentage points [25]. Puelma Touzel et al. [32] and Ghafouri et al. [11] observe that information flow complexity and uncertainty communication affect trust calibration. The EU AI Act [9] Articles 50 and 52 mandate machine-readable provenance metadata for AI-generated content, while the Digital Services Act (DSA) [7] requires platforms to implement content moderation transparency. The taxonomy in [24] shows LLMs achieve near-human mimicry (detection scores 0.46–0.50), crossing the "Indistinguishability Threshold" formalized in [20]. The C2PA standard [3] enables decentralized provenance verification, but most tools remain server-dependent. Building on the technical framework introduced in [27], Origin Lens provides a privacy-first, client-side alternative.

Positioning. Existing C2PA verification tools—including Adobe's Content Authenticity web inspector and the open-source `c2patool` CLI—require either cloud infrastructure or desktop environments. Mobile solutions such as Truepic operate as cloud-mediated capture pipelines rather than user-facing verification tools. Origin Lens differs in three ways: (1) all cryptographic verification executes on-device with no server dependency, (2) it combines C2PA with heuristic and watermark signals in a defense-in-depth pipeline, and (3) it provides graded confidence indicators designed for non-expert end users at the point of consumption.

3 System Architecture

Origin Lens employs a hybrid mobile architecture designed for performance and memory safety (see Figure 1). The core logic is implemented in **Rust** [30] for its memory safety guarantees [16] when parsing complex binary formats, while the user interface is built with **Flutter** [12] for cross-platform mobile deployment.

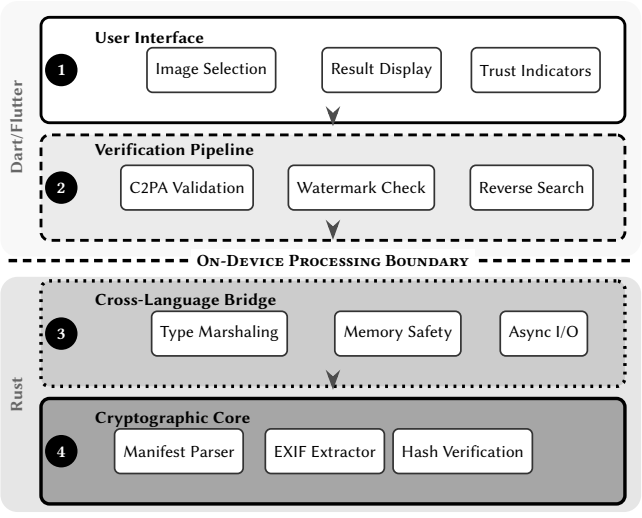


Figure 1: Origin Lens architecture: Flutter UI, verification pipeline, FFI bridge, and Rust core. Cryptographic operations execute on-device below the processing boundary.

Defense-in-Depth Strategy. Given epistemic uncertainty in modern media [11], we implement a four-layer pipeline: (1) *Cryptographic Provenance*—parse JUMBF boxes to validate C2PA manifests, verify X.509 trust chains, and compute SHA-256 hashes ensuring hard binding [3, 17, 43]; (2) *Heuristic Metadata*—analyze EXIF/IPTC for generative model artifacts (e.g., Stable Diffusion parameters); (3) *Watermark Detection*—detect imperceptible watermarks (e.g., SynthID) via API [4, 14, 15]; (4) *Contextual Verification*—opt-in reverse image search for prior attributions. Figure 2 shows the resulting decision workflow.

Privacy and Uncertainty Communication. In contrast to cloud-based detection services, Origin Lens processes C2PA and metadata entirely on-device via FFI bindings, following Privacy by Design principles [2] with privacy as the default. Cryptographic provenance and heuristic metadata analysis require no network access. Contextual verification (reverse image search) transmits image data to external services, leaving digital traces; this feature is therefore strictly opt-in and disabled by default, aligning with GDPR Article 25 data minimization [8, 44]. Communicating uncertainty is a known challenge [11]. Origin Lens implements a hierarchical confidence model: *high* (valid C2PA with trusted root), *medium* (EXIF patterns, watermarks), and *low* (opt-in reverse image search requiring user interpretation).

4 Security & Threat Modeling

We applied STRIDE threat modeling [36] to the Origin Lens verification pipeline, analyzing each architectural layer (Figure 1) for potential attack vectors.

Trust Model. Origin Lens relies on two trust anchors: (1) the C2PA trust list—root certificate authorities whose signatures are accepted as valid provenance—and (2) the integrity of the on-device binary. The system does not trust the network, the image source, or remote servers for core verification, reducing the attack surface to the local trust store and application package.

Identified Threats. At the *Verification Pipeline* layer, we identify three STRIDE-categorized risks: (1) *Manifest Stripping* (Tampering)—adversaries remove C2PA metadata during redistribution, a known limitation of content credentials [3]; (2) *Certificate Spoofing* (Spoofing)—attackers forge X.509 certificates to inject false provenance claims; and (3) *Denial of Service*—crafted manifests with deeply nested JUMBF boxes could cause excessive parsing time or memory exhaustion. At the *Cross-Language Bridge*, malformed input could trigger undefined behavior across the FFI boundary [31]; Rust's ownership model mitigates this by enforcing memory safety at compile time. The *User Interface* faces Information Disclosure risks if verification results are cached insecurely, and Repudiation risks if users cannot confirm verification correctness.

Mitigations. Origin Lens enforces hard binding validation [3]: SHA-256 hashes bind the manifest to image content—any modification invalidates the credential. Against spoofing, a local trust store validates certificate chains against known root authorities; we implement certificate pinning, enforce validity windows, and reject revoked certificates via embedded revocation lists. Against denial of service, the parser enforces depth and size limits on JUMBF structures with bounded memory allocation.

Open Challenges. The analog hole—screen capture or re-photographing a display—circumvents metadata-based provenance [15, 34]. Imperceptible watermarks (e.g., SynthID [14]) partially address this but

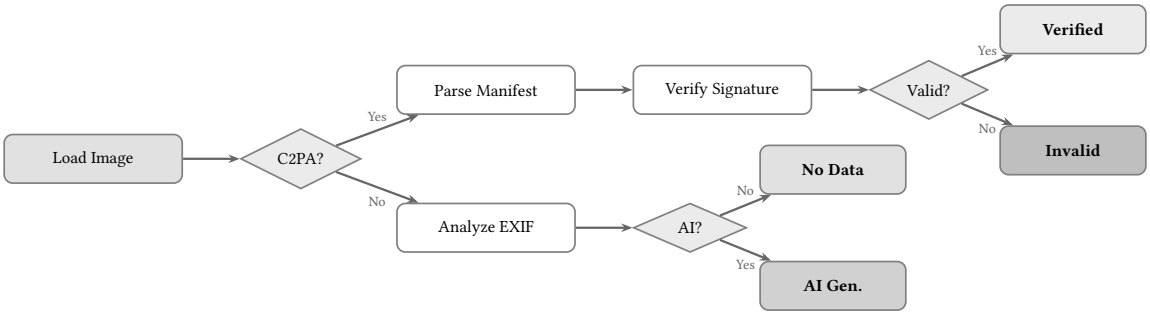


Figure 2: Analysis workflow: images with C2PA manifests undergo cryptographic verification; others fall back to EXIF-based heuristic detection.

depend on upstream embedding. We mitigate through defense-in-depth: heuristic metadata and watermark detection provide secondary signals when manifests are absent. Additionally, trust store freshness is critical; stale revocation data could allow compromised certificates to pass validation. Broad C2PA adoption remains a prerequisite for meaningful verification coverage [3].

5 Evaluation & Discussion

Performance. On iOS (iPhone 15 Pro), C2PA manifest validation completes in under 500 ms for 12 MP images, EXIF parsing in under 50 ms, and peak memory stays below 45 MB. Server-based verification requires a round trip of 200–800 ms plus image upload, whereas Origin Lens operates locally, avoiding both the latency and privacy cost of data transmission. The Rust core contributes to consistent performance: binary parsing runs without garbage-collection pauses, and the FFI bridge adds negligible overhead (< 1 ms per call). These results indicate that on-device provenance verification is practical for interactive use on current mobile hardware.

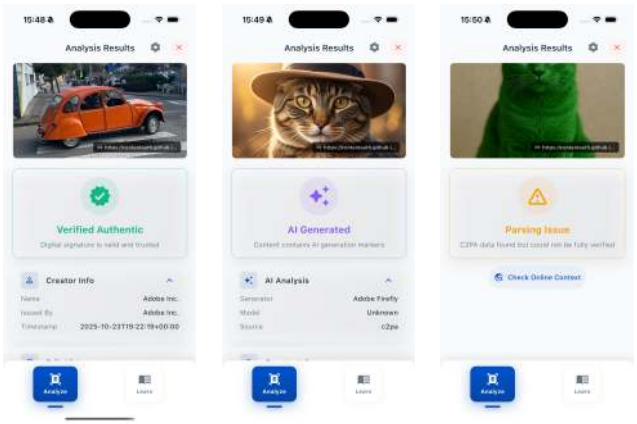
Result Communication. Origin Lens presents verification outcomes through a traffic-light-inspired UI [38, 42] with four states: *green* (valid C2PA from trusted root), *purple* (C2PA/EXIF indicating generative origin), *orange* (missing manifest or parse error), and *red* (hash mismatch or broken chain). This encoding was chosen because color-coded status displays reduce interpretation time compared to textual labels [42], and the traffic-light metaphor shows high comprehension in security-indicator studies [38].

Limitations. C2PA effectiveness depends on ecosystem adoption [3]. Adversarial actors may employ manifest stripping or analog-hole attacks. Heuristic detection faces distributional shift as models evolve [5, 28, 29, 37, 39, 40], and demographic predictors show weaker effects for AI-generated content [26]. The current evaluation covers a single device; broader benchmarking across device classes (mid-range Android, older iOS) and formats (HEIF, WebP, RAW) would strengthen generalizability. Critically, the present work evaluates technical performance but does not include a user study measuring whether participants correctly interpret verification signals or how these affect trust calibration and sharing behavior. Such a study is planned as the next evaluation phase, building on the dual-axis perception methodology developed in our JudgeGPT research [22]. Ultimately, technical tools alone are insufficient; verification and data literacy remain necessary for societal impact [19, 33]. Complementary approaches such as source-level credibility datasets [23] can augment image-level provenance with domain-level trust signals.

Data Availability. The Origin Lens source code is available at https://github.com/aloth/origin-lens under GPLv3. Related research datasets are archived on Zenodo: RogueGPT stimulus corpus (DOI: 10.5281/zenodo.18703138), JudgeGPT perception data (DOI: 10.5281/zenodo.18703385), and the CRED-1 domain credibility dataset (DOI: 10.5281/zenodo.19355308).

Regulatory Alignment. Origin Lens aligns with EU regulations that increasingly affect Web governance. Table 1 maps relevant requirements to framework features.

User Interface. Figure 3 shows Origin Lens in three verification scenarios. Table 2 details the four verification states communicated through a traffic-light-inspired visual language.



(a) Verified (b) AI Generated (c) Warning

Figure 3: Origin Lens verification results: (a) valid C2PA manifest with trusted chain, (b) AI-generated content detected via metadata, (c) parsing issue with incomplete chain.

6 Conclusion and Future Directions

As generative AI increases the volume of synthetic content on the Web, verification tools that operate independently of centralized platforms become relevant. Origin Lens shows that on-device,

Table 1: Regulatory alignment of Origin Lens features.

Regulation	Requirement	Alignment	User Benefit
EU AI Act	Art. 50, 52: Machine-readable provenance for AI content	C2PA parsing; AI metadata detection	Verify AI origin at point of consumption
GDPR	Art. 25: Privacy by Design; Art. 5: Data minimization	On-device processing; no server transmission	No personal data leaves the device
DSA [7]	Art. 14, 17: Moderation transparency	Independent client-side verification	User agency beyond platform controls
Cyber Resilience Act [10]	Security-by-design for digital products	Rust memory safety; X.509 validation	Reduced memory corruption risk

Table 2: Verification status indicators.

Status	Color	Condition
Verified	Green	Valid C2PA manifest with trusted certificate chain
AI Generated	Purple	AI origin detected via C2PA or EXIF markers
Warning	Orange	Missing manifest, expired certificate, or parse error
Invalid	Red	Hash mismatch or broken certificate chain

privacy-first image provenance verification is feasible, combining cryptographic standards (C2PA) with heuristic detection and graded uncertainty communication. The framework provides a client-side complement to platform-level governance, placing verification capability with the user rather than requiring trust in opaque moderation systems. Treating provenance verification as user-side infrastructure rather than a platform feature is consistent with Web Science research on decentralized trust and user agency.

Several directions remain open. First, lightweight on-device neural networks could enable pixel-level manipulation detection, extending coverage beyond metadata-based signals. Second, privacy-preserving federated aggregation of verification outcomes across users could detect emerging manipulation campaigns without centralizing raw image data. Third, as C2PA adoption grows, cross-jurisdictional alignment of provenance standards will be necessary to ensure interoperability across regulatory regimes. Finally, controlled user studies are needed to measure how on-device verification affects trust calibration, sharing behavior, and information literacy in ecologically valid Web browsing contexts. As our research continues, we invite experts to participate in our survey on verification practices: https://github.com/aloth/verification-crisis.

References

[1] Hannah Bloch-Wehba. 2022. Content Moderation as Surveillance. *Berkeley Technology Law Journal* 36, 3 (2022), 1297–1340. doi:10.15779/Z389C6S202

[2] Ann Cavoukian. 2010. Privacy by Design: The 7 Foundational Principles. Information and Privacy Commissioner of Ontario. Foundational framework for GDPR Article 25 implementation.

[3] Coalition for Content Provenance and Authenticity. 2025. *C2PA Technical Specification v2.3.* Technical Report. C2PA. https://spec.c2pa.org/specifications/specifications/2.3/specs/C2PA_Specification.html

[4] I.J. Cox, J. Kilian, F.T. Leighton, et al. 1997. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing* 6, 12 (1997), 1673–1687. doi:10.1109/83.650120

[5] Davide Cozzolino and Luisa Verdoliva. 2020. Noiseprint: A CNN-Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security* 15 (2020), 144–159. doi:10.1109/TIFS.2019.2916364

[6] Taylor Lynn Curtis, Maximilian Puelma Touzel, William Garneau, et al. 2025. Veracity: An Open-Source AI Fact-Checking System. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25).*

[7] European Parliament and Council of the European Union. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). *Official Journal of the European Union* L 277 (Oct. 2022), 1–102. http://data.europa.eu/eli/reg/2022/2065/oj

[8] European Parliament and Council. 2016. *Regulation (EU) 2016/679 (General Data Protection Regulation).* Regulation. Official Journal of the European Union. Article 25: Data protection by design and by default.

[9] European Parliament and Council. 2024. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act).* Regulation. Official Journal of the European Union. Articles 50, 52 on transparency and AI-generated content labeling.

[10] European Parliament and Council. 2024. *Regulation (EU) 2024/2847 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act).* Regulation. Official Journal of the European Union. http://data.europa.eu/eli/reg/2024/2847/oj Entered into force 10 Dec. 2024; applies from 11 Dec. 2027.

[11] Bijean Ghafouri, Shahrad Mohammadzadeh, Kellin Pelrine, et al. 2024. Epistemic Integrity in Large Language Models. arXiv:2411.06528

[12] Google. 2024. Flutter: Build apps for any screen. https://flutter.dev. Cross-platform UI toolkit.

[13] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society* 7, 1 (2020).

[14] Sven Gowal et al. 2025. SynthID-Image: Image watermarking at internet scale. arXiv:2510.09263 [cs.CR] https://arxiv.org/abs/2510.09263

[15] Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong. 2025. Watermarking and Detection of AI-Generated Images: A Survey. *arXiv preprint arXiv:2505.07894* (2025).

[16] Ralf Jung, Jacques-Henri Jourdan, Robbert Krebbers, et al. 2017. RustBelt: Securing the Foundations of the Rust Programming Language. *Proc. ACM Program. Lang.* 2, POPL, Article 66, 34 pages. doi:10.1145/3158154

[17] Daniel Kang, Tatsunori Hashimoto, Ion Stoica, et al. 2022. ZK-IMG: Attested Images via Zero-Knowledge Proofs to Fight Disinformation. arXiv:2211.04775 [cs.CR] https://arxiv.org/abs/2211.04775

[18] Victor Livernoche, Akshatha Arodi, Andreea Musulan, et al. 2025. OpenFake: An Open Dataset and Platform Toward Real-World Deepfake Detection. *arXiv preprint arXiv:2509.09495* (2025).

[19] Alexander Loth. 2021. *Decisively Digital: From Creating a Culture to Designing Strategy.* John Wiley & Sons, Inc., Hoboken, NJ, USA. https://www.wiley.com/en-us/Decisively+Digital%3A+From+Creating+a+Culture+to+Designing+Strategy-p-9781119737285

[20] Alexander Loth. 2026. The Indistinguishability Threshold: Measuring Cognitive Vulnerabilities to AI-Generated Disinformation. In *18th ACM Web Science Conference (WebSci Companion '26), May 26–29, 2026, Braunschweig, Germany, PhD Symposium.* ACM, New York, NY, USA. doi:10.1145/3795513.3807421

[21] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2024. Blessing or Curse? A Survey on the Impact of Generative AI on Fake News. doi:10.48550/arXiv.2404.03021 arXiv:2404.03021 [cs.CL]

[22] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Can Humans Tell? A Dual-Axis Study of Human Perception of LLM-Generated News. In *18th ACM Web Science Conference (WebSci Companion '26), May 26–29, 2026, Braunschweig, Germany.* ACM, New York, NY, USA. doi:10.1145/3795513.3807431

[23] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. CRED-1: An Open Multi-Signal Domain Credibility Dataset for Automated Pre-Bunking of Online Misinformation. (2026). doi:10.2139/ssrn.6448466 Preprint available at SSRN.

[24] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Eroding the Truth-Default: A Causal Analysis of Human Susceptibility to Foundation Model Hallucinations and Disinformation in the Wild. In *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)* (Dubai, United Arab Emirates). ACM, New York, NY, USA. To appear.

[25] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. Industrialized Deception: The Collateral Effects of LLM-Generated Misinformation on Digital Ecosystems. In *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)* (Dubai, United Arab Emirates). ACM, New York, NY, USA. To appear.

[26] Alexander Loth, Martin Kappes, and Marc-Oliver Pahl. 2026. The Verification Crisis: Expert Perceptions of GenAI Disinformation and the Case for Reproducible Provenance. In *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)* (Dubai, United Arab Emirates). ACM, New York, NY, USA. To appear.

[27] Alexander Loth, Dominique Conceicao Rosario, Peter Ebinger, Martin Kappes, and Marc-Oliver Pahl. 2026. Origin Lens: A Privacy-First Mobile Framework for Cryptographic Image Provenance and AI Detection. In *Companion Proceedings of the ACM Web Conference 2026 (WWW '26 Companion)*. ACM, New York, NY, USA. Also available as arXiv:2602.03423.

[28] J. Lukas, J. Fridrich, and M. Goljan. 2006. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security* 1, 2 (2006), 205–214. doi:10.1109/TIFS.2006.873602

[29] Hannes Mareen and Dante et al. Vanden Bussche. 2023. Comprint: Image Forgery Detection and Localization Using Compression Fingerprints. In *Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges*, Jean-Jacques Rousseau and Bill Kapralos (Eds.). Springer Nature Switzerland, Cham, 281–299.

[30] Nicholas D Matsakis and Felix S Klock. 2014. The Rust Language. *ACM SIGAda Ada Letters* 34, 3 (2014), 103–104.

[31] Ian McCormack, Joshua Sunshine, and Jonathan Aldrich. 2025. A Study of Undefined Behavior Across Foreign Function Boundaries in Rust Libraries. In *Proceedings of the 2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*. IEEE, Ottawa, ON, Canada. doi:10.1109/ICSE55347.2025.00167

[32] Maximilian Puelma Touzel, Sneheel Sarangi, Gayatri Krishnakumar, et al. 2025. Simulating public discourse in digital societies by giving social media to multimodal AI agents. In *Proceedings of the IJCAI Demo Track 2025*.

[33] Sijia Qian, Cuihua Shen, and Jingwen Zhang. 2022. Fighting cheapfakes: using a digital media literacy intervention to motivate reverse search of out-of-context visual misinformation. *Journal of Computer-Mediated Communication* 28, 1 (11 2022), zmac024. doi:10.1093/jcmc/zmac024 arXiv:https://academic.oup.com/jcmc/article-pdf/28/1/zmac024/47194917/zmac024.pdf

[34] Bhaktipriya Radharapu and Harish Krishna. 2024. RealSeal: Revolutionizing Media Authentication with Real-Time Realism Scoring. In *Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24)*. ACM, New York, NY, USA. doi:10.1145/3678957.3678960

[35] Catherine Régis, Florian Martin-Bariteau, Jake Okechukwu Effoduh, et al. 2025. *AI in the Ballot Box: Four Actions to Safeguard Election Integrity and Uphold Democracy*. Technical Report. IVADO Global Policy Briefs. https://doi.org/10.32920/28382087

[36] Adam Shostack. 2014. Threat Modeling: Designing for Security. *Wiley* (2014). STRIDE threat modeling methodology.

[37] Saud Sohail, Syed Muhammad Sajjad, Adeel Zafar, et al. 2025. Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning. *Information* 16, 4 (2025). doi:10.3390/info16040270

[38] Borce Stojkovski, Gabriele Lenzini, and Vincent Koenig. 2021. "I Personally Relate It to the Traffic Light": A User Study on Security & Privacy Indicators in a Secure Email System Committed to Privacy by Default. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing (SAC '21)*. ACM, New York, NY, USA, 1–10. doi:10.1145/3412841.3441998

[39] Camille Thibault, Jacob-Junqi Tian, Gabrielle Péloquin-Skulski, et al. 2025. A Guide to Misinformation Detection Data and Evaluation. *arXiv preprint arXiv:2411.05060* (2025).

[40] Luisa Verdoliva. 2020. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing* 14, 5 (2020), 910–932. doi:10.1109/JSTSP.2020.3002101

[41] Sze Yuh Nina Wang, Samantha C. Phillips, Kathleen M. Carley, et al. 2025. Limited effectiveness of psychological inoculation against misinformation in a social media feed. *PNAS Nexus* 4, 6 (2025), pgaf172.

[42] Christopher D. Wickens and Anthony D. Andre. 1990. Proximity Compatibility and Information Display: Effects of Color, Space, and Objectness on Information Integration. In *Human Factors*, Vol. 32. SAGE Publications, 61–77. Traffic light metaphor for status displays.

[43] Mingyang Xie, Manav Kulshrestha, Shaojie Wang, et al. 2022. PROVES: Establishing Image Provenance using Semantic Signatures. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 3017–3026. doi:10.1109/WACV51458.2022.00307

[44] Yunkang Yang, Trevor Davis, and Matthew Hindman. 2023. Visual misinformation on Facebook. *Journal of Communication* 73, 4 (02 2023), 316–328. doi:10.1093/joc/jqac051 arXiv:https://academic.oup.com/joc/article-pdf/73/4/316/50973350/jqac051.pdf