

How Is Artificial Intelligence Reshaping Visual Trust? A Review of Photography, Image Manipulation, and the Crisis of Authenticity in Digital Media

Myrah Ratnani

Student, Dhirubhai Ambani International School, Mumbai, India

Abstract

Photography has always been privileged in its position as evidence, being regarded as the image that "must have been there" because the image itself is a record of a physical meeting between a scene and a surface. However, that relationship becomes more complicated when it comes to artificial intelligence. Today, generative systems can also create photorealistic images, without relying on a photographed referent, and the more and more widely available editing tools can manipulate authentic images at a level of semantics. This review explores the implications of that change on visual trust. It takes as its theoretical moorings three vital concepts: Barthes's idea of photographic indexicality, the recorded history of photographic manipulation, and Chesney and Citron's idea of the liar's dividend. The following foundations are challenged with the latest evidence of human detection, machine detection and provenance technologies. Human performance becomes less reliable and less consistent: Nightingale and Farid found that for images, participants were only 48.2% accurate when distinguishing real from AI-synthesized faces, and a multimodal study in 2025 showed a 49.4% accuracy for images. While machine detectors are successful in highly controlled benchmarks, cross-generator generalization, compression, resizing and model evolution are ongoing problems. The central issue no longer just has to be whether a specific image is bogus, the review says. The more fundamental issue is that the question of convincing undermines the presumption of all images, even of a real object, and particularly the presumption of their verisimilitude. Detection must not therefore be the primary concern for authentication. New systems like C2PA Content Credentials and SynthID move the issue away from the pixel to the origin and transformations, but don't validate semantic verity. The paper argues that visual trust is transforming from an indexical to a chain of custody model (provenance, forensic analysis, contextual verification, institutional accountability).

Keywords: generative AI; visual trust; photography; deepfakes; image forensics; provenance; C2PA; synthetic media

1. Introduction

Photographs have been the subject of a different treatment in the history of mass media compared to paintings, illustrations, and verbal testimony. A photo could be framed selectively, staged, cropped, retouched or otherwise deceptive captioned, but still seemed to capture a casual trace of something

standing before a camera. Roland Barthes spoke about this relation in the light of the photographic referent and the “that-has-been” – the unique assertion that the thing shown had once been there in front of the camera (Barthes, 1981). Susan Sontag, also, while asserting that each picture is a selection and an interpretation, claimed that photography is unusually bound to the physical world. This inherited evidentiary contract is interrupted by generative AI. With the introduction of diffusion models, it's now possible to generate highly realistic scenes from text or image prompts without any external event that corresponds. This isn't just an upgraded version of Photoshop. It is an ontological change in the image. A synthetic image like a photo can have the visual grammar of photography without the traditional causal relationship between the image and a real referent. The question posed in this review is how is AI transforming visual trust and how can a blend of detection, provenance and institutional verification maintain the evidential value of digital images? The argument takes a two-part course. The first is that it sets the current age of AI apart from the longer history of image manipulation. Secondly, it analyzes the practical methods in use today for separating real, altered and synthetic images. The main idea is that visual trust is shifting from Pixel trust to Provenance trust. The question in the future isn't “Does this image look real?” but, “What verifiable chain of evidence supports this image as real, and how it has been transformed, what context it was created in, and who publishes it?”

2. Method and Analytical Framework

This paper is a structured review of literature. It brings together basic theory in photography, a history of the media, peer-reviewed research in human perception, computer vision, synthetic media detection and digital provenance along with primary technical documentation from standards bodies and technology developers. The review is structured in the light of the three pillars of the research design: photographic indexicality, the historical continuity of manipulation, and the liar's dividend. It compares these pillars to three domains of contemporary evidence: human detection, machine detection under distribution shift, and systems for provenance or watermarking. There are four types of visual trust which are clearly distinguished. Causal trust is whether or not the scene depicted had a physical referent. Process trust asks 'how was this image taken and manipulated? Semantic trust is asking if the image, caption, and the surrounding claim is a true representation of the event. Institutional trust is about whether a credible actor will state the picture and its chain of custody. Normally, these categories were collapsed into one intuition, namely, if the image was photographic, it was treated as evidence, as is traditional photography. Generative AI is the one that divides them.

3. From Indexicality to Synthetic Referents

Barthes's theory provides a way of understanding how photographs gained an evidentiary status in the absence of being a neutral record. Meaning has always been mediated by lens choice, exposure, cropping, darkroom processing, sequencing and captions. But there was one indexical constraint that remained: there was some light from some external scene that was involved in the production of the photograph. That causal relationship did not ensure that the photograph was true, but it ensured that it was truly different from a drawing of an event that never happened. The history of manipulation has proved that indexicality was no guarantee of the veracity of photographs. Stalin's manipulation of images is documented by David King in the instances of the removal of individuals like Nikolai Yezhov from historical photographs after political purges (King, 1997). The 1982 cover of the National Geographic that featured the Pyramids of Giza moved closer together to fit into a vertical composition

was an early public controversy over electronic image alteration (Goldberg, 2016). These examples are important because they stop one from getting caught up in an overly nostalgic argument that photographs were once perfectly truthful. The difference is that AI does not manipulate but its production function. In the past, it took an original picture, expertise, time, and limited edits to be able to falsify an earlier image. Generative systems minimize all four constraints. They can generate the whole context, semantically change identities and objects, generate numerous variants with very little additional cost, and tailor the output with natural-language prompts. Diffusion models learn the statistical structure of image data and then reverse a noise process; latent diffusion makes the process more efficient by running most of the denoising in a learned latent space and by allowing for flexible conditioning with text due to the use of cross attention (Ho, et al., 2020; Rombach, et al., 2022). This transition can be termed as "manipulation of evidence" to "simulation of evidence. A conventional edited photograph starts with the referent and alters the representation of the photograph. A completely synthetic image can start from scratch. This disparity is the basis for today's modern authenticity crisis.

Table 1: Three Technological Regimes of Visual Trust

Regime	Relation to physical referent	Typical manipulation constraint	Scale	Primary trust problem
Analog photography	Physical capture required	Manual retouching, staging, cropping	Relatively limited	Misleading representation of a real capture
Digital editing	Physical capture usually present	Software-based alteration	High	Untraceable modification of captured evidence
Generative AI	Physical capture not required	Prompted synthesis and semantic editing	Extremely high	Photographic appearance without photographic occurrence

Source: Author's synthesis from the literature reviewed in this paper.

4. Human Detection: The Collapse of "Looking Carefully"

A natural reaction to synthetic imagery is the visual response: look at hands, reflections, text, shadows, anatomy or background consistency. This is still helpful as a heuristic, but not a secure authentication method. Technical innovation can be seen as a dying breed as the artifacts of one generation are shortened and refined in the next, and advanced users can "curate" out the "glaring failures". This instability is reflected in the empirical evidence. 315 participants were solicited by Nightingale and Farid (2022) to tell the difference between real faces and those generated by StyleGAN2. The mean accuracy was 48.2%, which is not significantly different from chance. Interestingly, the synthetic faces were also perceived as more believable than the real faces. Cooke et al. (2025) conducted a larger and more environmentally focused study with 1,276 participants of image, audio, video, and audiovisual media. The accuracy of overall detection was 51.2%, and for image, 49.4%. There was clear asymmetry in the participants' ability to recognize fully authentic media, compared with their ability to correctly identify

synthetic content, particularly in places where users believe that "real" = "authentic". Recent studies on the diffusion-generated photograph do not soothe these nerves, but provide additional nuance. In their study of 749,828 judgments from 50,444 participants, Kamali, et al., 2025 found that the complexity of the scene, type of artifact, viewing time, and human manipulation of synthetic images were important factors in detection. The implication here is key: there is no checklist of visual errors that can be used reliably to distinguish between synthetic and authentic content. Any human examination can raise suspicion, but the lack of visible artifacts cannot prove authenticity.

5. Machine Detection: High Benchmark Scores, Fragile Generalization

Automated detectors seem to present a more promising approach as they can take advantage of statistical signals that the human eye can't see. This was backed by early work. Wang et al. (2020) demonstrated a surprising ability to generalize from images of one convolutional generator to several different generators that were unseen during training. The paper's title, though, carried the word "for now"—meaning the underlying issue: Detectors are able to learn the "fingerprint" of the generator, which is not permanent and can change as synthesis architectures change. The shift from GANs to diffusion models revealed that. Ojha et al (2023) demonstrated that traditional real/fake classifiers trained from old generators tend to perform poorly on new families of generators. Their work recast the problem as out-of-distribution generalization, instead of the usual classification. The GenImage benchmark was also created with the new addition of cross-generator and degraded-image tasks, which were added because a high score on the original generator does not necessarily translate into a good score after compression, blur, resize, or switching to a new model (Zhu, et al., 2023). There are newer methods that have been developed to enhance generalization. FakeInversion is based on inversion capabilities of a pre-trained diffusion model and shows better results on out-of-sample generators like DALL-E 3 (Cazenavette, et al., 2024). There are also improvements in robustness against unseen architectures and "laundered" images, which are post-processed (Cozzolino, et al., 2024). However, this advancement does not remove the inherent disbalance in generation and detection. An open but evolving set of synthetic images must be recognized by a detector, but a generator need only generate images that are not recognized by the detector at this moment. Detection should therefore be considered in a probabilistic manner for this reason. Evidence does not equal an authenticated chain of origin. Photographs that are truly of a scene can be discredited by false positives, while photographs of real objects can be certified by false negatives. Those two mistakes can be institutional risks in high stakes settings.

6. The Liar's Dividend: When Authentic Images Also Lose Value

Chesney and Citron (2019) characterize another side of the deepfakes, the liar's dividend. When convincing fabrication is generally recognized as a possibility, individuals who are caught by authentic evidence can obtain a new defense: the evidence itself can now be called into question as fake. This makes the authenticity issue a symmetrical issue. An image may be taken as a true image and a real image may be treated as a false image. The findings of the experiments indicate that uncertainty can be as significant as outright deception. Vaccari and Chadwick (2020) discovered that people were more likely to be uncertain than to be completely convinced of the political deepfake and that such uncertainty had a negative effect on trust in news on social media. It is epistemologically important, since institutions like journalism, courts, science, and elections rely not just on detecting lies, but on keeping a reasonable margin of error where they can be believed if they are true. The "liar's dividend" thus shifts the purpose

of authentication. A system that is successful will not simply be able to label synthetic images once they have been generated. It also has to include affirming evidence of authentic content. Otherwise, the very act of skepticism can be weaponised. Perhaps the most significant impact of generative AI is not on the number of mis.created images, but the extra amount of verification on every true image.

7. From Detection to Provenance

That's why there is a technical shift from detection to provenance. Content Credentials are cryptographically verifiable records, which describe the origin and transformation history of an asset, defined by the Coalition for Content Provenance and Authenticity (C2PA). C2PA 2.4 distinguishes between "hard" bindings, which are cryptographic hashes linked to an individual asset, and "soft" bindings, which include perceptual fingerprints or invisible watermarks that can assist in recovering provenance when metadata is removed from the file. It also enables signed assertions and trust lists to be used for verifying the identity of the credential issuer (C2PA, 2026). The idea is to shift the question from "Is this fake?" to "Who created this asset, on what device or software, and what transformations have since been made?" That is a more robust evidentiary architecture which considers authenticity a chain-of-custody issue. Google DeepMind's SynthID is an example of a watermarking system, which handles a similar failure mode. SynthID will add a watermark to synthetic content that is not detectable even after common content manipulations like cropping, filtering, other color changes, and lossy compression. Google makes clear, however, that watermarking isn't a foolproof solution. Content created by systems that are not part of the system can also lack watermarks, and genuine images created using other systems might not include any provenance metadata (Google DeepMind, 2023). First and foremost, C2PA's tagging is not synonymous of veracity. A cryptographically valid credential can be used to assert that a particular signer made a particular assertion, and that the record has not been altered, but it cannot prove that the scene was staged, the capture was accurate, or that the caption was semantically true. Similarly, the lack of a credential does not prove that an image is not authentic. The provenances remedies the origin problem more effectively than the meaning problem.

Table 2: What Different Trust Technologies Can and Cannot Establish

Method	Strongest contribution	Central limitation
Human inspection	Rapid contextual skepticism	Near-chance performance for many modern synthetic images
Forensic detector	Statistical evidence of synthesis or manipulation	Model shift, laundering, false positives and false negatives
Invisible watermark	Generator-linked machine-readable signal	Requires participating generator; robustness is not absolute
C2PA provenance	Signed chain of origin and edits	Does not establish semantic truth; adoption is incomplete
Institutional verification	Context, source corroboration, editorial or legal accountability	Slower, resource-intensive, dependent on trusted institutions

Source: Author's synthesis from detection and provenance literature.

8. A Layered Model of Visual Trust

The evidence indicates that there is no one solution that returns to the old adage of “seeing is believing.” Layered means a more defensible model. The first is provenance: wherever possible, capture devices, generation systems and editing software should generate signed records. The second is forensic analysis: Detectors are still useful for unsigned or non-standard content, especially as probabilistic indicators, not as simple yes/no questions. The third is contextual verification, which may not be tested with cryptography: reverse-image search, geolocation, timestamp analysis, source corroboration, and comparison with independent evidence. The fourth – Institutional accountability – Publishers, courts, platforms, and news organizations require clear guidelines on when visual evidence is accepted, marked, disputed, or saved. This makes a new definition of authenticity. Just because no detector finds evidence of the synthesis, then the image is not necessarily considered trustworthy. When and where independent signals converge, you should feel more trusting – if the line of trust can be traced back to a credible source, the pixels can be co-ordinated and technically coherent, the context is consistent and the source is accountable. When they are conflicting, however, uncertainty should be represented, not hidden. The larger implication is a change in visual literacy. Critical viewing in the twentieth century entailed asking questions of author and subject, what was left out and the effect of captions on interpretation. For the generative era, those questions are still important, but not enough. The viewer also has to wonder if the scene was ever actually shot.

9. Limitations

This is not to say that every photograph looks like an AI-generated image, nor that every automatic detector is faulty. This depends on the generator, the content, its curation, compression, detector architecture, and evaluation protocol. The characteristics of human detection studies range widely as well in stimulus design and ecological realism. Provenance standards are currently being developed, and will take effect over a longer period in the future depending on the uptake of cameras, model providers, editing tools, platforms and institutions. Lastly, there is visual authenticity and there is visual authenticity. Even a truly authentic photograph can be posed, cropped or miscaptioned.

10. Conclusion

The use of AI has turned the idea of visual trust upside down by challenging one assumption that has long been in play within the photographic record, namely that an image like a photograph must have a photographed event. History shows that photographs never were truthful by themselves, and AI capabilities have made the act of falsification more scalable, costlier, faster, and referential. It is not ecologically feasible to make a perceptual solution. Human detectors can achieve almost random accuracy on realistic synthetic imagery, while machine detectors are susceptible to changes in the environment and evolution of the model. Simultaneously the liar's dividend argues that the problem cannot be solved by simply identifying fakes; there also needs to be stronger "plus" evidence for the authentic images. So it's a move from plausible vision to verifiable source that is most promising. C2PA, watermarking, forensic detection, and institutional verification provide solutions to different aspects of the problem of trust and should not be seen as mutually exclusive. Visual evidence is unlikely to return to the unquestioned photographic authority. It is a more explicit system of evidence in which the trust is built on a chain of origin, technical integrity, contextual corroboration and accountable human institutions.

11. References

1. Barthes R., “Camera Lucida: Reflections on Photography”, Hill and Wang, 1981.
2. Cazenavette G., Sud A., Leung T., Usman B., “FakeInversion: Learning to Detect Images from Unseen Text-to-Image Models by Inverting Stable Diffusion”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, 10759–10769. <https://doi.org/10.1109/CVPR52733.2024.01023>
3. Chesney R., Citron D., “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security”, California Law Review, 2019, 107 (6), 1753–1820. <https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security/>
4. Coalition for Content Provenance and Authenticity (C2PA), “Content Credentials: C2PA Technical Specification, Version 2.4”, 2026. https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html
5. Cooke D., Edwards A., Barkoff S., Kelly K., “As Good as a Coin Toss: Human Detection of AI-Generated Content”, Communications of the ACM, 2025, 68 (10), 100–109. <https://doi.org/10.1145/3729417>
6. Cozzolino D., Poggi G., Corvi R., Nießner M., Verdoliva L., “Raising the Bar of AI-generated Image Detection with CLIP”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2024, 4356–4366. <https://doi.org/10.1109/CVPRW63382.2024.00439>
7. Goldberg S., “How We Spot Altered Pictures”, National Geographic, 2016. <https://www.nationalgeographic.com/magazine/article/editors-note-images-and-ethics>
8. Google DeepMind, “Identifying AI-generated images with SynthID”, 2023. <https://deepmind.google/blog/identifying-ai-generated-images-with-synthid/>
9. Ho J., Jain A., Abbeel P., “Denoising Diffusion Probabilistic Models”, Advances in Neural Information Processing Systems, 2020, 33. <https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html>
10. Kamali N., Nakamura K., Kumar A., Chatzimpampas A., Hullman J., Groh M., “Characterizing Photorealism and Artifacts in Diffusion Model-Generated Images”, Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, Article 794, 1–26. <https://doi.org/10.1145/3706598.3713962>
11. King D., “The Commissar Vanishes: The Falsification of Photographs and Art in Stalin’s Russia”, Metropolitan Books, 1997.
12. Nightingale S.J., Farid H., “AI-synthesized faces are indistinguishable from real faces and more trustworthy”, Proceedings of the National Academy of Sciences, 2022, 119 (8), e2120481119. <https://doi.org/10.1073/pnas.2120481119>
13. Ojha U., Li Y., Lee Y.J., “Towards Universal Fake Image Detectors that Generalize Across Generative Models”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 24480–24489. <https://doi.org/10.1109/CVPR52729.2023.02345>
14. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B., “High-Resolution Image Synthesis with Latent Diffusion Models”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 10684–10695. <https://doi.org/10.1109/CVPR52688.2022.01042>
15. Sontag S., “On Photography”, Farrar, Straus and Giroux, 1977.

16. Vaccari C., Chadwick A., “Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News”, *Social Media + Society*, 2020, 6 (1). <https://doi.org/10.1177/2056305120903408>
17. Wang S.-Y., Wang O., Zhang R., Owens A., Efros A.A., “CNN-Generated Images Are Surprisingly Easy to Spot... for Now”, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 8695–8704. <https://doi.org/10.1109/CVPR42600.2020.00872>
18. Zhu M., Chen H., Yan Q., Huang X., Lin G., Li W., Tu Z., Hu H., Hu J., Wang Y., “GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image”, *Advances in Neural Information Processing Systems*, 2023, 36. <https://doi.org/10.52202/075280-3398>