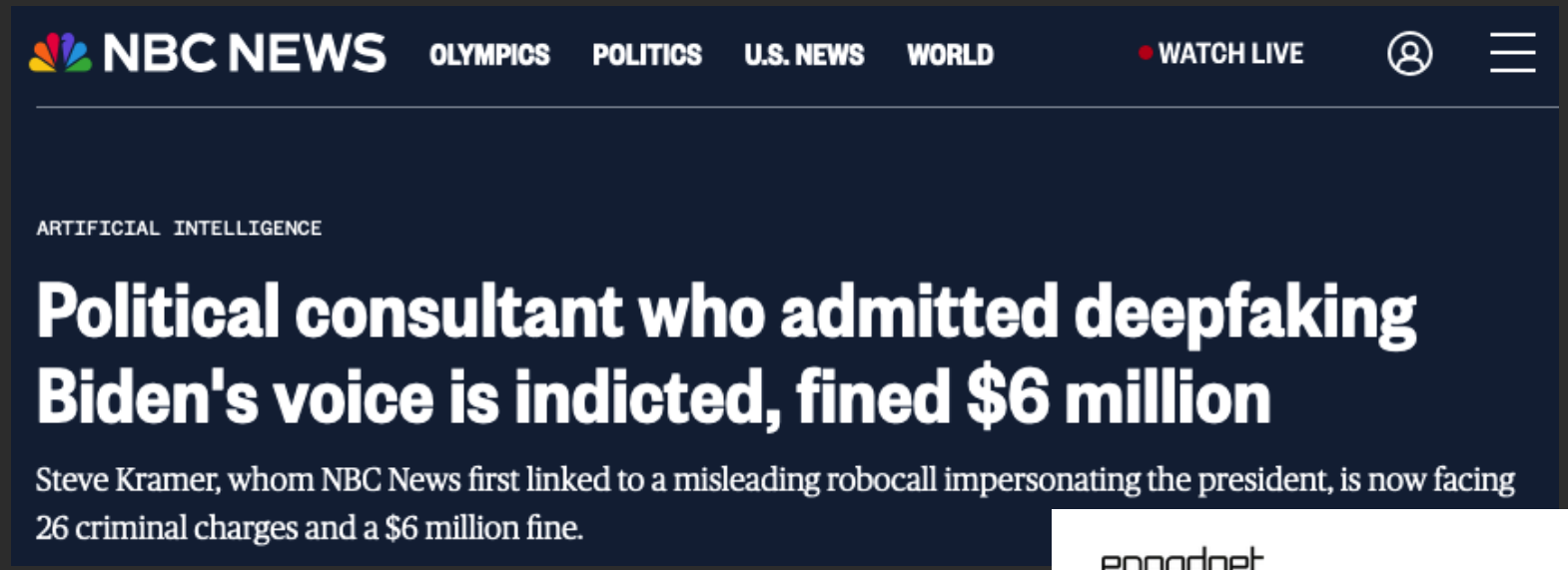
Tracing Origins: Navigating Content Authenticity in the Deepfake Era

Speaker(s):
Peleus Uhley
Principal Scientist, Adobe Inc.
Co-Chair Threats and Harms Working Group
Coalition for Content Provenance and Authenticity (C2PA)

Why?

Deepfakes have been used in the upcoming US election



NBC NEWS OLYMPICS POLITICS U.S. NEWS WORLD WATCH LIVE

ARTIFICIAL INTELLIGENCE

Political consultant who admitted deepfaking Biden's voice is indicted, fined $6 million

Steve Kramer, whom NBC News first linked to a misleading robocall impersonating the president, is now facing 26 criminal charges and a $6 million fine.

engadget

Elon Musk shared a doctored Harris campaign video on X without labeling it as fake

The video uses a clone of the VP's voice to make it seem as though she's saying disparaging things about herself and Joe Biden.

Cheyenne MacDonald · Weekend Editor

Sun, July 28, 2024 at 10:26 AM PDT · 1 min read

1

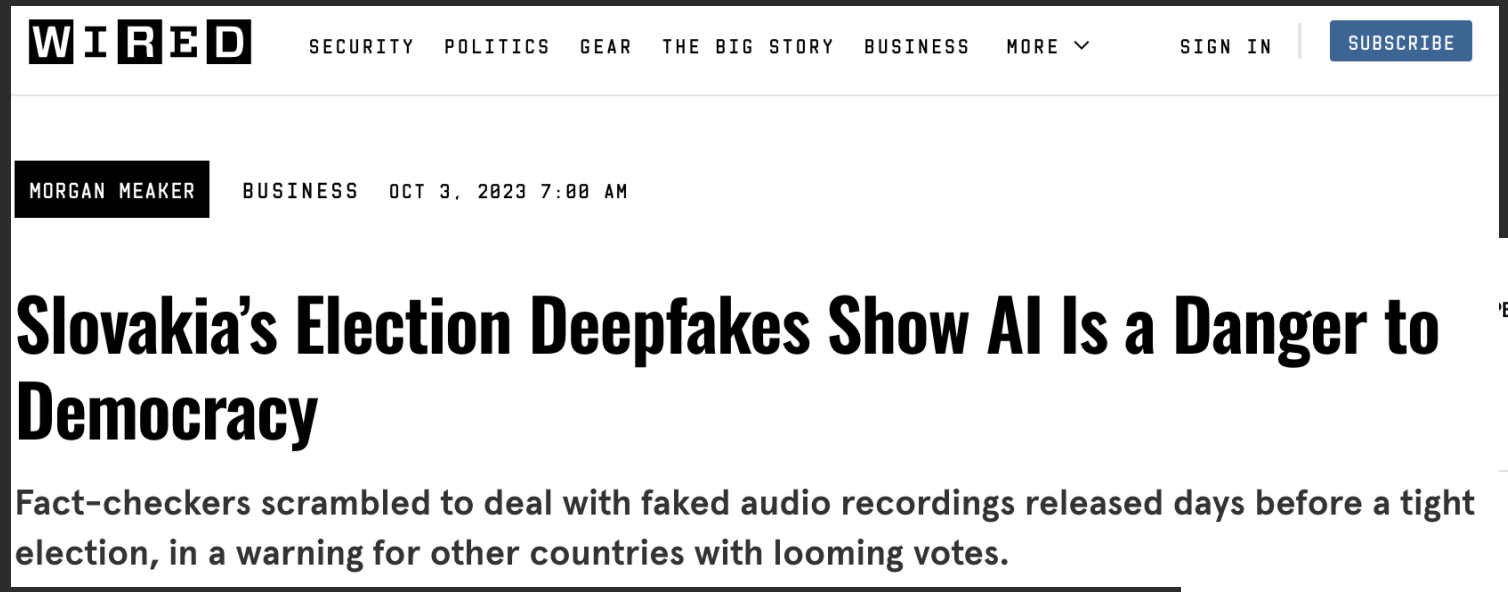
As spotted by *The New York Times*, Elon Musk shared an altered version of Kamala Harris' campaign video on Friday night that uses a deepfake voiceover

TIME

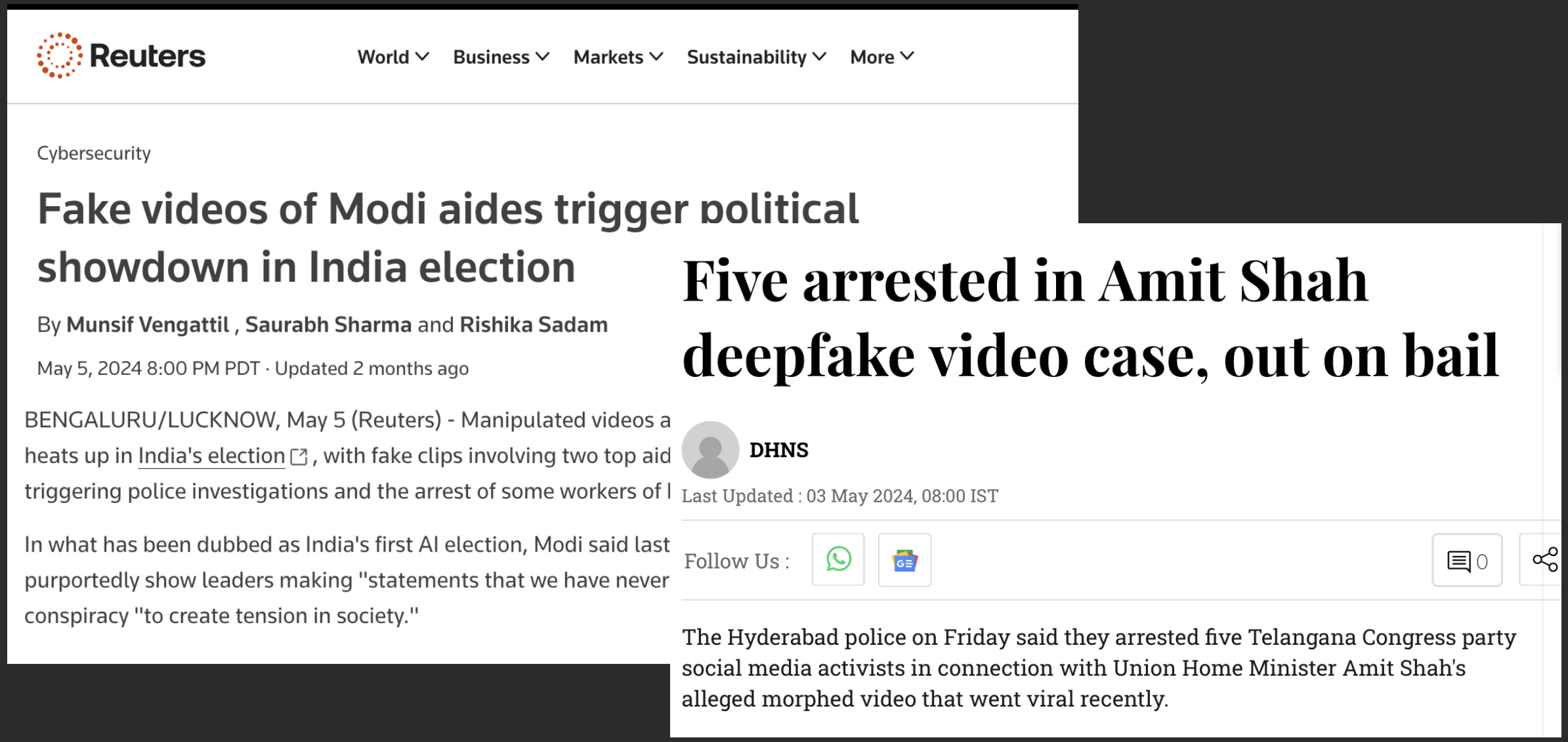
2024 is not just *an* election year. It's perhaps *the* election year.

Globally, more voters than ever in history will head to the polls as at least 64 countries (plus the European Union)—representing a combined population of about 49% of the people in the world—are meant to hold national elections, the results of which, for many, will prove consequential for years to come.

Deepfakes are a global issue



Deepfakes are a global issue (cont.)



Reuters

World Business Markets Sustainability More

Cybersecurity

Fake videos of Modi aides trigger political showdown in India election

By **Munsif Vengattil , Saurabh Sharma** and **Rishika Sadam**

May 5, 2024 8:00 PM PDT · Updated 2 months ago

BENGALURU/LUCKNOW, May 5 (Reuters) - Manipulated videos a
heats up in India's election , with fake clips involving two top aid
triggering police investigations and the arrest of some workers of I

In what has been dubbed as India's first AI election, Modi said last
purportedly show leaders making "statements that we have never
conspiracy "to create tension in society."

Five arrested in Amit Shah deepfake video case, out on bail

DHNS

Last Updated : 03 May 2024, 08:00 IST

Follow Us :

0

The Hyderabad police on Friday said they arrested five Telangana Congress party social media activists in connection with Union Home Minister Amit Shah's alleged morphed video that went viral recently.

Split perspectives

WIRED

NILESH CHRISTOPHER VARSHA BANSAL THE BIG STORY MAY 20, 2024 6:00 AM

Indian Voters Are Being Bombarded With Millions of Deepfakes. Political Candidates Approve

India's elections are a glimpse of the AI-driven future of democracy. Politicians are using audio and video deepfakes of themselves to reach voters—who may have no idea they've been talking to a clone.

Senators Coons, Blackburn, Klobuchar, Tillis introduce bill to protect individuals' voices and likenesses from AI-generated replicas

Formal bill introduction follows October 2023 introduction of discussion draft

JULY 31, 2024

- Moody's warned in a report on Wednesday that AI-generated deepfake political content could contribute to the election integrity threats and "undermine the credibility of U.S. institutions."
- A recently proposed FCC rule would require political TV, video and radio ads to disclose if they used AI-generated content, part of a push against deep fakes and other manipulated content, but not covering social media.
- While companies like Meta and Google are self-policing AI-manipulated content, industry standards are lacking.

Tech companies stepping into the fight

Adobe amazon ANTHROPIC arm IIElevenLabs

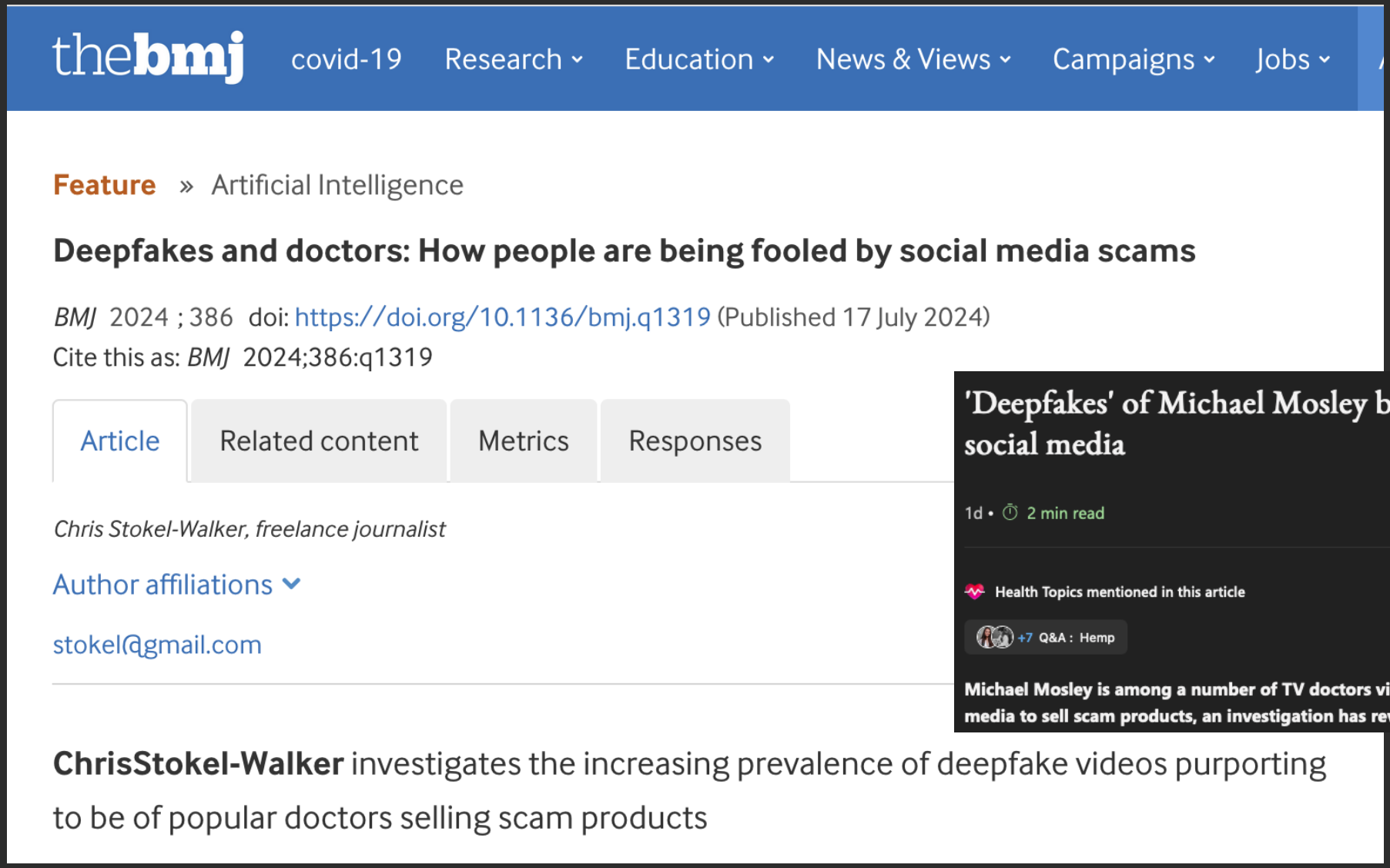
Gen GitHub Google IBM Inflection INTUIT

LG AI Research LinkedIn McAfee Microsoft Meta

NetApp NOTA OpenAI SNAPCHAT stability.ai

TikTok TREND MICRO TrueMedia.Org truepic X

Four out of five AI bots approve



Generating mass confusion

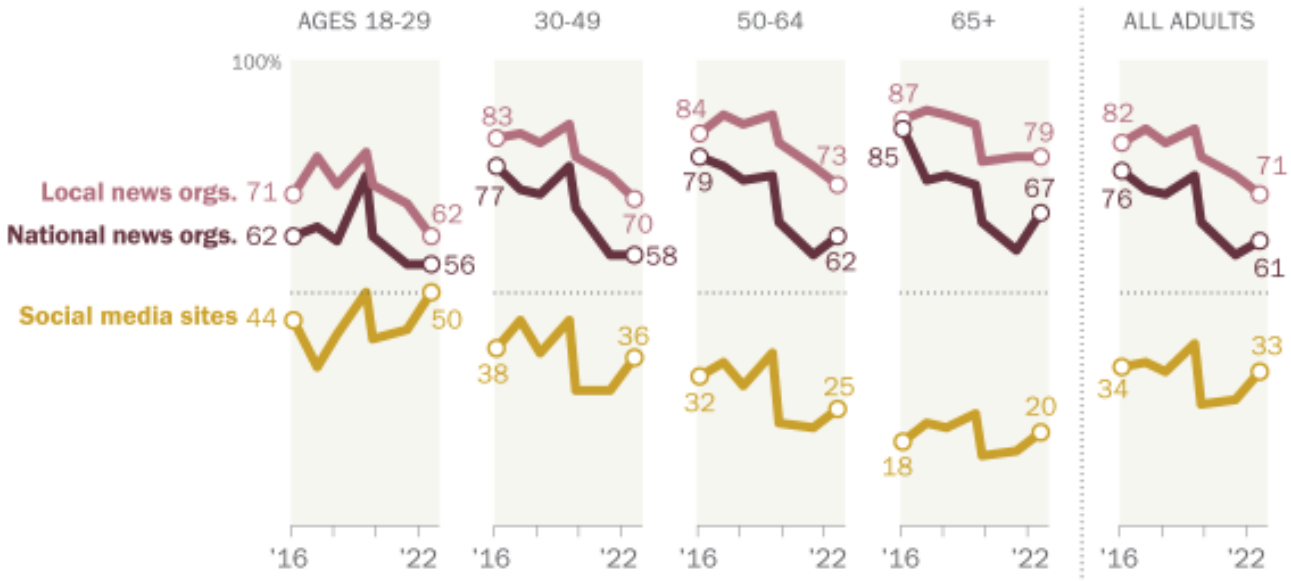
- An AI-generated scene that is the product of AI hallucination—made up information that may seem plausible but is not true—that depicts an explosion near the Pentagon was shared around the internet in May 2023, causing general confusion and turmoil on the stock market. [14]
- A deepfake video showed Ukrainian President Volodomyr Zelenskyy telling his country to surrender to Russia. [15] More recently, several Russian TV channels and radio stations were hacked and a purported deepfake video of Russian President Vladimir Putin was aired claiming he was enacting martial law due to Ukrainians invading Russia. [16]

SHORT READ | OCT 27, 2022

U.S. adults under 30 now trust information from social media almost as much as from national news outlets

Half of 18- to 29-year-olds say they have at least some trust in the information they get from social media sites.

*% of U.S. adults who say they have **some or a lot of trust** in the information they get from ...*



Social media is a prominent channel for announcements

Instagram

Log In Sign Up

JOSEPH R. BIDEN, JR.

joebiden · Follow

July 21, 2024

My Fellow Americans,

Over the past three and a half years, we have made great progress as a Nation.

Today, America has the strongest economy in the world. We've made historic investments in rebuilding our Nation, in lowering prescription drug costs for seniors, and in expanding affordable health care to a record number of Americans. We've provided critically needed care to a million veterans exposed to toxic substances. Passed the first gun safety law in 30 years. Appointed the first African American woman to the Supreme Court. And passed the most significant climate legislation in the history of the world. America has never been better positioned to lead than we are today.

I know none of this could have been done without you, the American people. Together, we overcame a once in a century pandemic and the worst economic crisis since the Great Depression. We've protected and preserved our Democracy. And we've revitalized and strengthened our alliances around the world.

It has been the greatest honor of my life to serve as your President. And while it has been my intention to seek reelection, I believe it is in the best interest of my party and the country for me to stand down and to focus solely on fulfilling my duties as President for the remainder of my term.

I will speak to the Nation later this week in more detail about my decision.

For now, let me express my deepest gratitude to all those who have worked so hard to see me reelected. I want to thank Vice President Kamala Harris for being an extraordinary partner in all this work. And let me express my heartfelt appreciation to the American people for the faith and trust you have placed in me.

I believe today what I always have: that there is nothing America can't do – when we do it together. We just have to remember we are the United States of America.

https://joebiden.com

Joe and Kamala are on board -- are you? Chip in today to power our campaign.

$25 $46 $100 $250 $1000 OTHER

If you've saved your payment information with ActBlue Express, your donation will go through immediately.

BIDEN HARRIS

ACCOMPLISHMENTS ISSUES TAKE ACTION STORE DONATE

TOGETHER, WE CAN

FINISH THE JOB

FOR THE

AMERICAN PEOPLE!

EdSmart data collected from Pew studies (2023)

What impact is it having on everyday Americans?

- In North America, the proportion of deepfakes more than doubled from 2022 to Q1 2023. This proportion jumped from 0.2% to 2.6% in the U.S. (Business Wire)
- 52% of Americans have changed the way they use social media (Pew)
- 43% have lessened their overall news intake (Pew)
- 15% of Americans encounter altered videos or images often (Pew)
- 33% of Americans say they hardly ever or never come across altered videos or images (Pew)
- 66% of Americans say they at least sometimes come across altered videos and images that are intended to mislead (Pew)
- 63% say videos and images altered to mislead the public cause a great deal of confusion (Pew)
- Deepfakes are becoming more common in financial scams, with a 300% increase in deepfake-based fraud attempts reported in 2020. (Source: Digital Guardian)

EdSmart data (cont.)

How does Deepfake affect America's political atmosphere?

- 68% of US adults say made-up news and information greatly impacts Americans' confidence in government institutions (Pew)
- 54% of US adults say made-up news and information is having a major impact on our confidence in each other (Pew)
- 73% see a lot of made-up news and information being generated around two major topics: politics and elections (Pew)
- 63% of Americans say made-up or altered videos and images create a great deal of confusion about the facts of current issues and events (Pew)
- 77% of both Republicans and Republican-leaning independents and Democrats and Democratic leaners favor restrictions on altered videos over protecting the freedom to publish and access them (Pew)

Who's responsible for solving this problem?

- 53% of Americans believe journalists have the greatest responsibility to reduce made-up news (Pew)
- 12% say the responsibility falls on the government (Pew)
- 9% say tech companies are responsible (Pew)

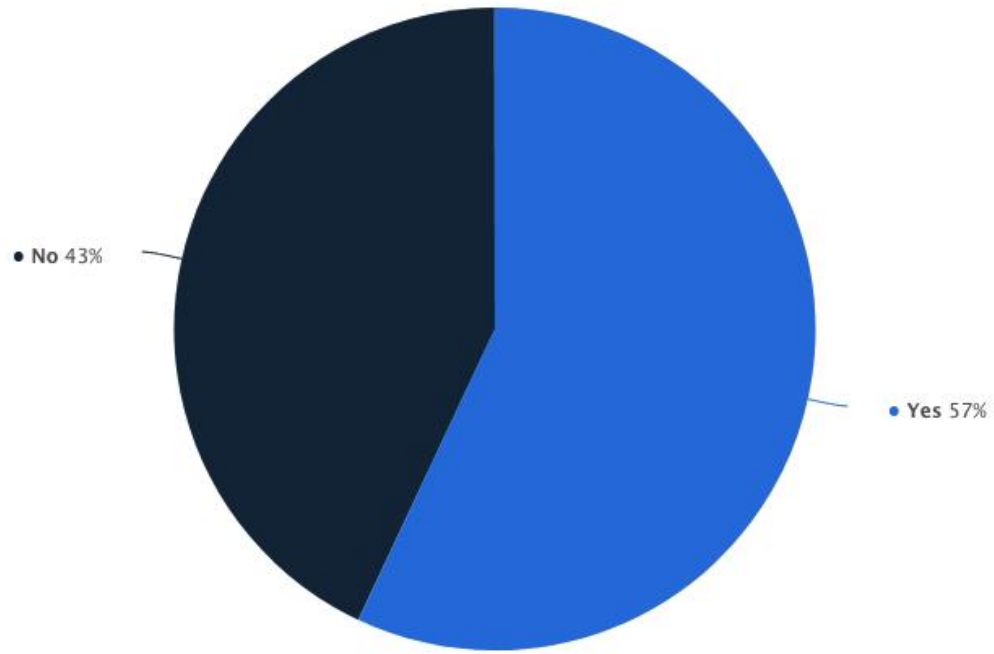
statista

Search Statistics

Prices & Access | Statistics | Reports | Insights | Research AI NEW | Daily Data | Services | Login

Technology & Telecommunications › Software

Share of consumers who say they could detect a deepfake video worldwide as of 2022



Additional Information

© Statista 2024

Show source

DOWNLOAD

PDF + | XLS + | PNG + | PPT +

Source

→ Show sources information

→ Show publisher information

→ Use Ask Statista Research Service

Release date

August 2022

Region

Worldwide

Survey time period

2022

Number of respondents

16,000 respondents

Supplementary notes

Original question: "Do you think you would be able to tell the difference between a real video and a

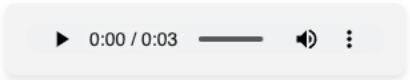
Deepfakes are a problem across media types

News

by Hany Farid, UC Berkeley Professor, CAI Advisor
July 22, 2024

In collaboration with Sarah Barrington, a Ph.D. student at UC Berkeley, we have launched a new study to determine just how realistic these AI-generated voices are and whether a cloned voice sounds like the original speaker's voice.

In this study, participants listen to a set of voices (one at a time), half of which are real and half of which are AI-generated. Although we are still collecting data, we have completed a pilot study with 50 participants who each listened to a total of 40 short voice recordings. The average accuracy on this task was 65%, slightly better than chance performance of 50%. There was only a small bias in which accuracy for real voices was 68% and accuracy for fake voices was 62%. In other words, participants were slightly more likely to say a recording was real. You can test yourself on a set of 16 voices to see how you do.



An audio clip from a collection of recordings consisting of natural human and AI-generated voices.

We also asked participants how they thought they were distinguishing the real from the fake. We received some interesting insights, including:

- The person was breathing or taking breaths between words.
- Fake had too much enunciation.
- Fake had no speaking errors.

While these preliminary results suggest that AI-generated voices are passing through the uncanny valley, they do not mean that all AI-generated voices are indistinguishable from reality. The snippets of voices that participants heard were relatively short, between 3 and 10 seconds, and did not feature yelling, laughing, or anything that reflected strong emotions. If, however, generative AI continues along its current trajectory, it seems likely that sooner or later it is going to be very difficult to perceptually distinguish the real from the fake.

So how good are deepfakes today?

So how good are deepfakes today?

It doesn’t matter.

So how good are deepfakes today?

It doesn’t matter.

Tomorrow they will be better.

Who?

Coalition for Content Provenance and Authenticity (C2PA)

A better view of the C2PA member list

- ***Big Tech***: Microsoft, Google, Adobe, AWS, EA, Sony, Fastly, Akamai, etc.
- ***Chipset Manufacturers***: Intel, ARM, Qualcomm, etc.
- ***Camera Manufacturers and Apps***: Leica, Nikon, Canon, Fujufilm, Truepic, etc.
- ***News and Media Orgs***: BBC, NY Times, RIAA, Warner Bros, Universal, CBC, France TV, WDR, etc.
- ***CAs and Identity Providers/Orgs***: DigiCert, CyberTrust, Identity, Cheqd, ISCC Foundation, Privateid, Toothpic, etc.
- ***AI Groups***: OpenAI, Animechain.ai, ElevenLabs, etc.
- ***Non-profits***: Witness, Partnership on AI, The Society Library, etc.

C2PA background

- Founded in February 2021 and now has almost 100 members
- The Coalition for Content Provenance and Authenticity (C2PA) is a project of the Linux Foundation Joint Development Foundation, a Washington-based 501c6 non-profit, that brings together the efforts of the Content Authenticity Initiative (CAI) and Project Origin.
- Mission is to develop technical specifications that can establish content provenance and authenticity at scale to give publishers, creators, and consumers the ability to trace the origin of media.
- Currently on version 2.X of the specification (around 200 pages)
- Covers multiple media types (images, videos, audio files, PDFs, etc.)
- Liaisons with ISO, IPTC, ETSI, PDF Association

C2PA creates both the specification and informative documents

Technical Specifications

- Content Credentials
- Attestations

Guidance & Informative Documents

- Explainer
- Guidance for Implementers
- User Experience Guidance
- Security Considerations
- Harms Modelling
- Guidance for Artificial Intelligence and Machine Learning

PDF Versions of these documents are also available via the Download button in the page header.

Where?

Camera implementations

PetaPixel Reviews Guides Learn Glossary Podcast Newsletters

Nikon Made a New In-Camera ‘Digital Watermark’ to Go on Top of C2PA

JAN 09, 2024 JARON SCHNEIDER

Leica

Nick Rains explains: Content Credentials in the M11-P

Nick Rains is a professional photographer and principal Instructor at Leica Akademie Australia. In his educational video, he shares an in-depth perspective on the newest technology built into the Leica M11-P.

PetaPixel Reviews Guides Learn Glossary Podcast Newsletters

Canon and Reuters Develop New Photo Authentication Technology

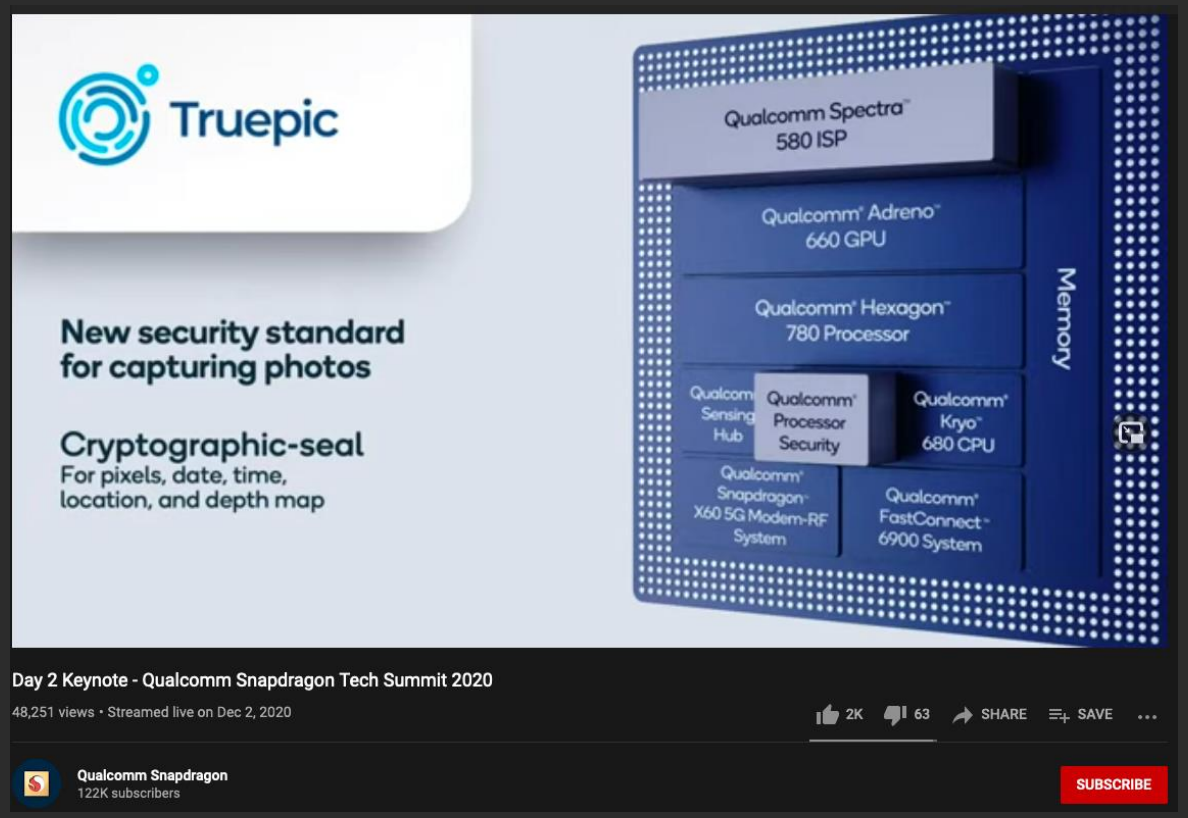
AUG 31, 2023 JEREMY GRAY

Canon

THOMSON REUTERS

Hardware-secured photo capture on mobile devices

- Qualcomm Processor Security confirms the device's identity with Qualcomm's wireless edge services on Qualcomm Snapdragon 3 chip.
- The Qualcomm Spectra 580 Image Signal Processor (ISP) is responsible for the secure image sensor data acquisition.
- This works with TruePic's mobile application and related services for creating the CAI signature.
- https://www.youtube.com/watch?v=XvvtrKg7Xuk



Online services

TIKTOK / CREATORS / TECH

TikTok is adding an ‘AI-generated’ label to watermarked third-party content

/ TikTok says it’s the first social media platform to support the Content Credentials metadata tags for AI-generated content.

OpenAI

Search for articles...

All Collections › Privacy and policies › Policy FAQ › C2PA in DALL·E 3

C2PA in DALL·E 3

C2PA standard, OpenAI's implementation, and C2PA metadata

Updated over a week ago

BBC

Sign in

Home News Sport Business Innovation Cu

Research & Development

Home About Projects Publications Blog Contact Us Careers

Mark the good stuff: Content provenance and the fight against disinformation

Posted by **Charlie Halford** on 5 Mar 2024, last updated 7 Mar 2024

Content Authenticity Initiative (CAI) - *https://contentauthenticity.org/*

- *The Content Authenticity Initiative is a cross-industry community of over 3,000 major media and technology companies, educational groups, non-profits, and government and policy organizations founded and led by Adobe in 2019.*

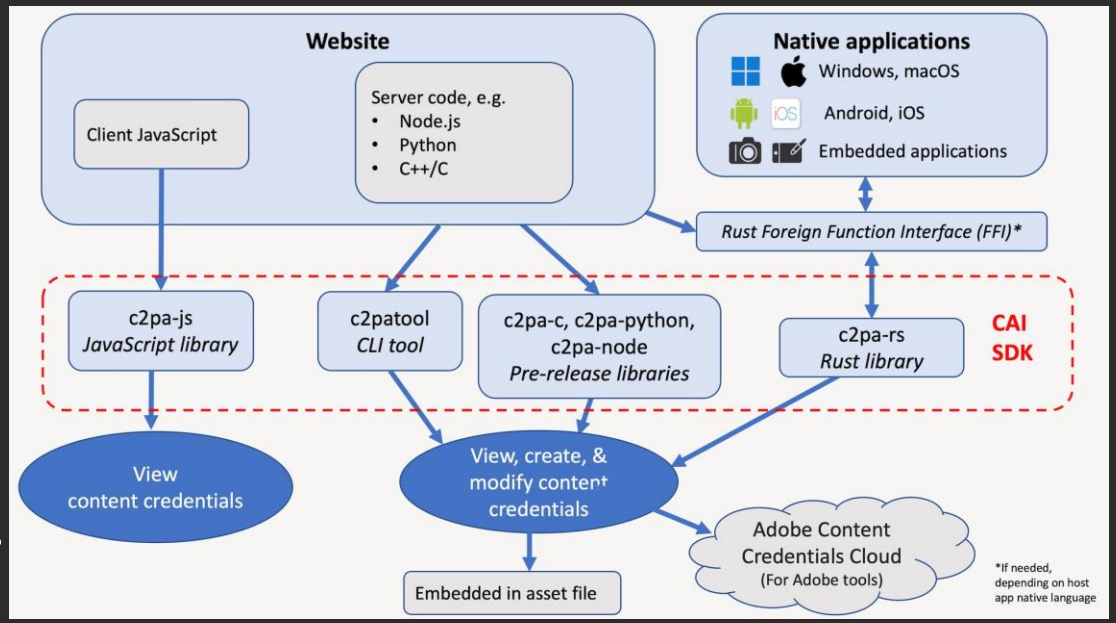
  The CAI open-source SDK (software development kit) is a set of tools and libraries that enable developers to create, verify, and display Content Credentials based on C2PA standards.. *The Content Authenticity Initiative media literacy curricula are crafted to prepare school-age students with critical media and visual literacy skills to help them better navigate the ever-changing digital information landscape.*

- *https://github.com/contentauth/*

Content Authenticity Initiative

Open-source implementations

- https://github.com/contentauth/ by the Content Authenticity Initiative (CAI)
  - C2patool – displays content credentials from images
  - Rust and JavaScript SDKs – SDKs to manage content-credentials
  - c2pa-attacks tool – Security tool to test UI and parsers.

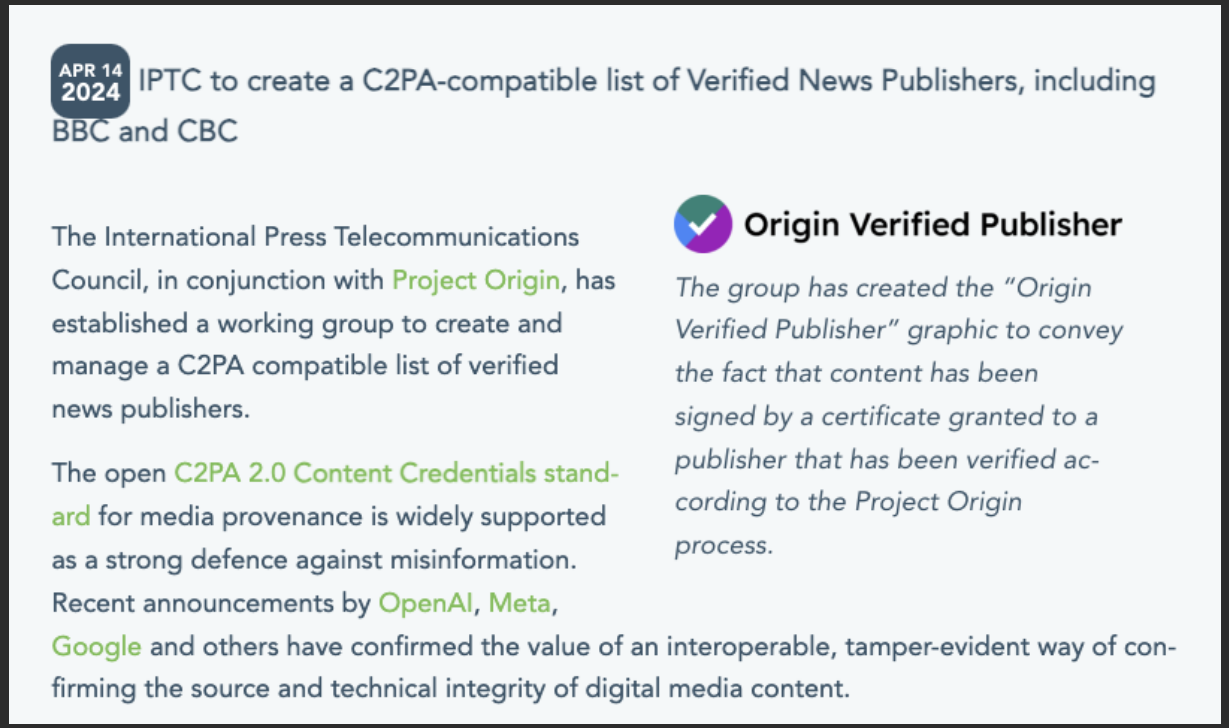


- Chrome extension by Digimarc based on the CAI JavaScript SDK
  - https://github.com/digimarc-corp/c2pa-content-credentials-extension

Project Origin - https://www.originproject.info/

Project Origin
Protecting Trusted Media

- Project Origin is “building a list of identity-verified news providers, to anchor media provenance technologies like C2PA, in conjunction with media education and synthetic media detection techniques will help to establish a foundation for trust in media.”
- An alliance of leading organizations from the publishing and technology worlds, working together to create a process where the provenance and technical integrity of content can be confirmed. Establishing a chain of trust from the publisher to the consumer
- Includes both publishing and consumer-side



APR 14 2024 IPTC to create a C2PA-compatible list of Verified News Publishers, including BBC and CBC

The International Press Telecommunications Council, in conjunction with Project Origin, has established a working group to create and manage a C2PA compatible list of verified news publishers.

The open C2PA 2.0 Content Credentials standard for media provenance is widely supported as a strong defence against misinformation. Recent announcements by OpenAI, Meta, Google and others have confirmed the value of an interoperable, tamper-evident way of confirming the source and technical integrity of digital media content.

Origin Verified Publisher

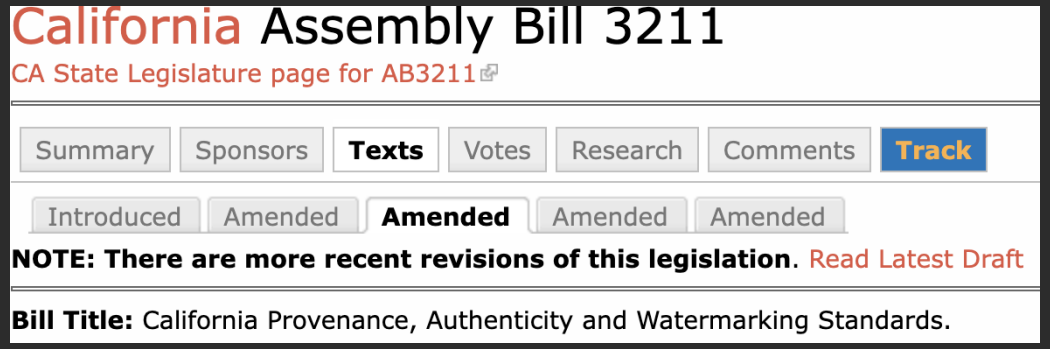
The group has created the “Origin Verified Publisher” graphic to convey the fact that content has been signed by a certificate granted to a publisher that has been verified according to the Project Origin process.

Creator Assertions Working Group (CAWG)

- Created in response to identity changes in version 2.0 of the specification
  - Identity is extremely complex to implement with complex laws surrounding it.
  - Separating identity from C2PA allows for choice and reduces complexity in the C2PA specification
- Creator Assertions Working Group (CAWG)
  - Not a part of the C2PA specification or the C2PA but it is compliant with the C2PA specification.
  - Conceptually adapted from Trust over IP (ToIP).
- https://creator-assertions.github.io/

Work with other standards organizations

- Joint Photographic Experts Group (JPEG) committee is working on JPEG Trust - ISO/DIS 21617-1
  - JPEG Trust defines a framework for establishing trust in media. This framework addresses aspects of authenticity, provenance and integrity through secure and reliable annotation of media assets throughout their life cycle. https://jpeg.org/jpegtrust/index.html
  - Will be published “soon” and is compatible with C2PA with added functionality
- C2PA is a liaison with ISO TC 171/SC 2 (authenticity of information) as Working Group 14
  - Planning to create an ISO version of the 2.X of the C2PA specification
- ISO helps C2PA with government requirements



What?

What does the specification/organization try to accomplish?

- Informs on the sources (tools, techniques, etc.) used in the creation of the content (Authenticity)
- Provides a history of the content with audit trails and timestamps (Provenance)
- Support asset formats across images, videos, audio, fonts, and documents
- Does not require cloud storage or distributed ledgers/blockchain but allows for it
- The organization recently received a $500,000 grant from the Societal Resilience Fund to launch an educational campaign aimed at providing clarity and awareness of the current landscape of digital disclosure methods and best practices.
- Provide indicators that you can use to decide whether or not to trust a piece of content

Levels of Trust

(0-100)

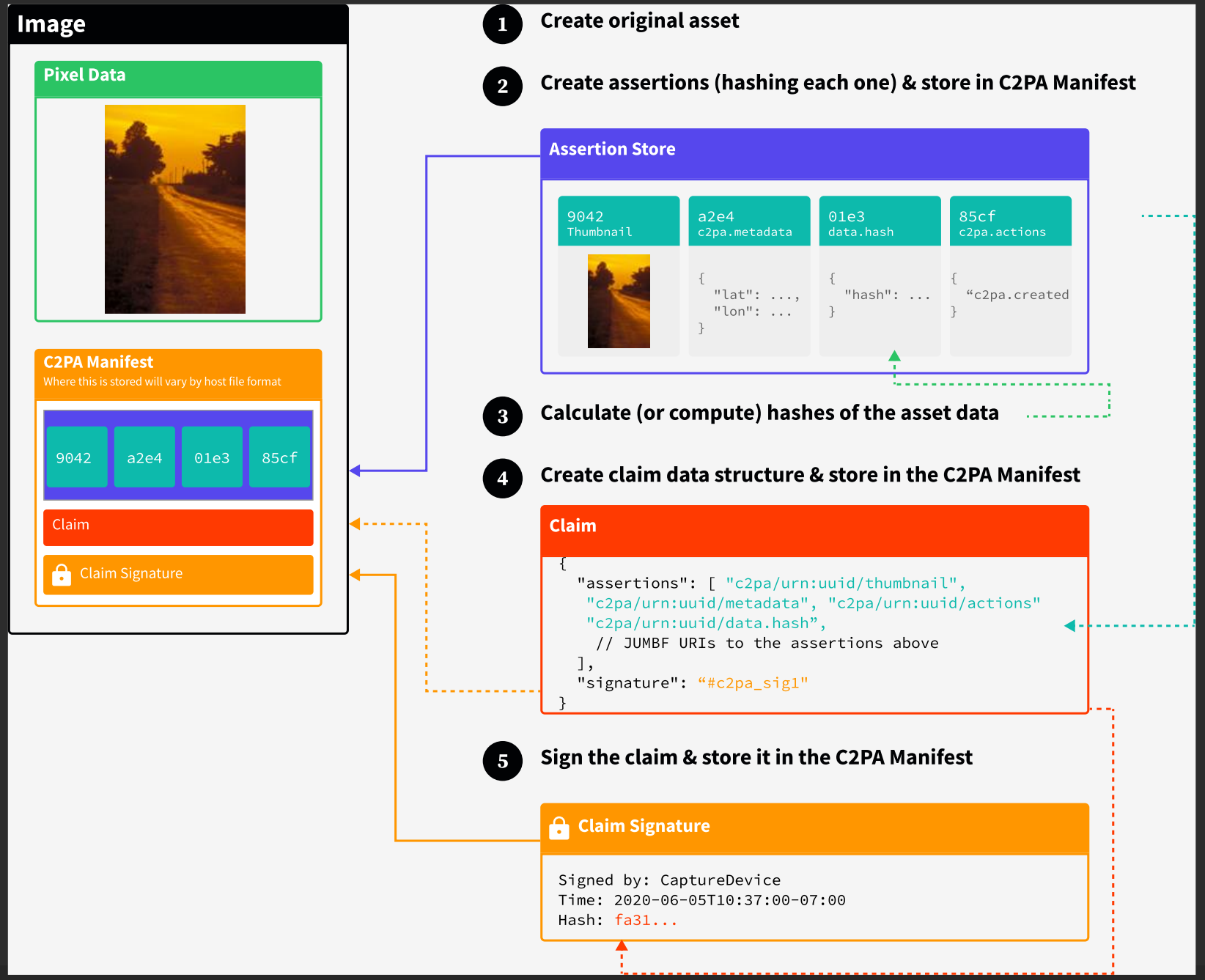
How we measure the trust of something we see

What are not goals of the specification/organization?

- It does not tell you whether to trust the content
  - How you interpret the data provided is up to you
  - Whether you trust the identities associated with the content is up to you
- It can’t guarantee the date the asset was created. It can only guarantee the date it was signed. *
  - * At the implementation level, some device apps will sign at the time of creation. However, that is implied by the tool and not the specification.
- C2PA does not try to solve detection. The group only focuses on provenance, education, and policy

How?

Creation



The implementation is based on existing specifications

3. Normative References

3.1. Core Formats

- CBOR
- JSON
- JSON-LD
- JPEG universal metadata box format (JUMBF)

3.2. Schemas

- CDDL
- JSON Schema
- Dublin Core Metadata Initiative

3.3. Digital & Electronic Signatures

- X.509 Certificates
- JSON Web Algorithms (JWA)
- CBOR Object Signing and Encryption (COSE)
- Using RSA Algorithms with COSE Messages
- Online Certificate Status Protocol (OCSP)
- Internet X.509 PKI Time-Stamp Protocol
- CBOR Object Signing and Encryption (COSE): Header Parameters for Carrying and Referencing X.509 Certificates
- Algorithms and Identifiers for the Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile
- Internet X.509 Public Key Infrastructure: Additional Algorithms and Identifiers for DSA and ECDSA
- Algorithm Identifiers for Ed25519, Ed448, X25519, and X448 for Use in the Internet X.509 Public Key

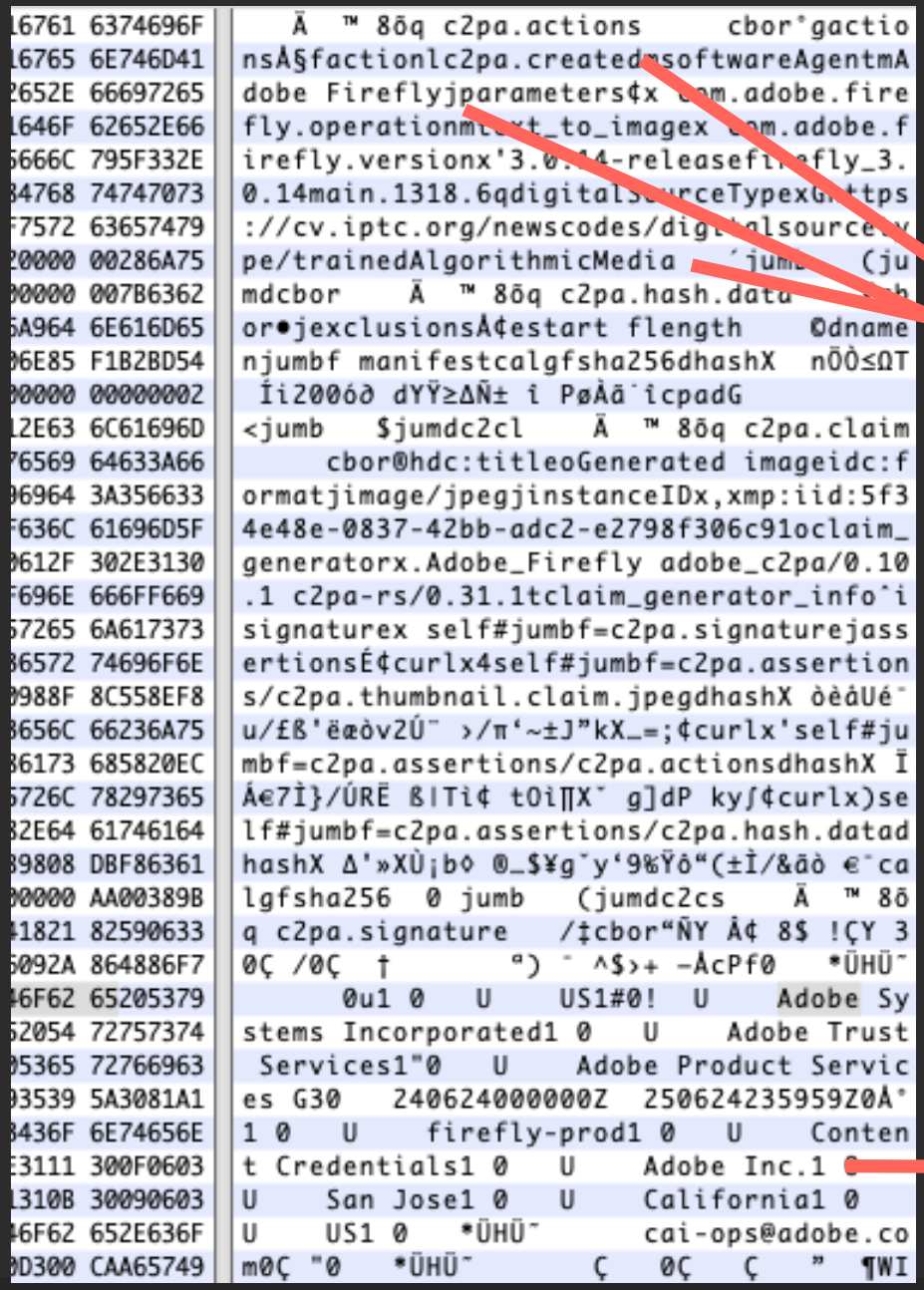
Cryptography

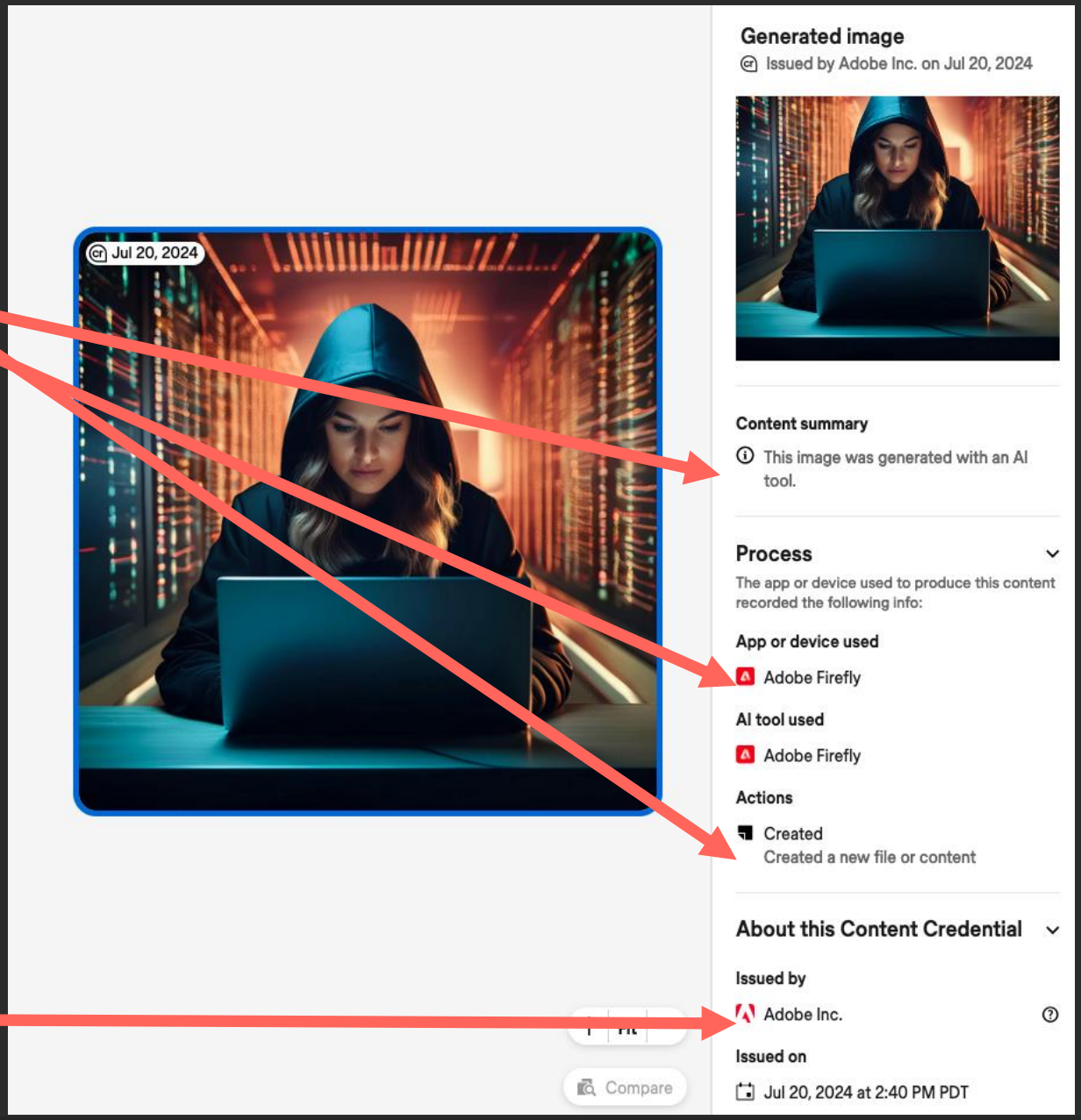
- There is a concept of a “trust list” but it is not the browser or OS trust list.
  - The list is not yet finalized.
  - CAs are involved in the discussions.
  - You can add your own CAs locally.
- The specification supports OCSP and OCSP stapling
  - I am currently in the process of writing revocation guidance.
- Support perceptual hashes (soft bindings) for similarity searches and traditional hashes (hard bindings)

C2PA and Generative AI

- Identifying assets and regions of interest that have been created/modified by AI
- Information about the “recipe” used in the creation (prompts, model, etc.)
- Allows creators to label “do not train” for their content
  - Compatible with European Union Text and Data Mining (TDM) standard
- International Press Telecommunications Council (IPTC) support
  - IPTC recommends that software creating images using trained AI algorithms uses the “Digital Source Type” value of “trainedAlgorithmicMedia”
  - https://cv.iptc.org/newscodes/digitalsourcetype/trainedAlgorithmicMedia

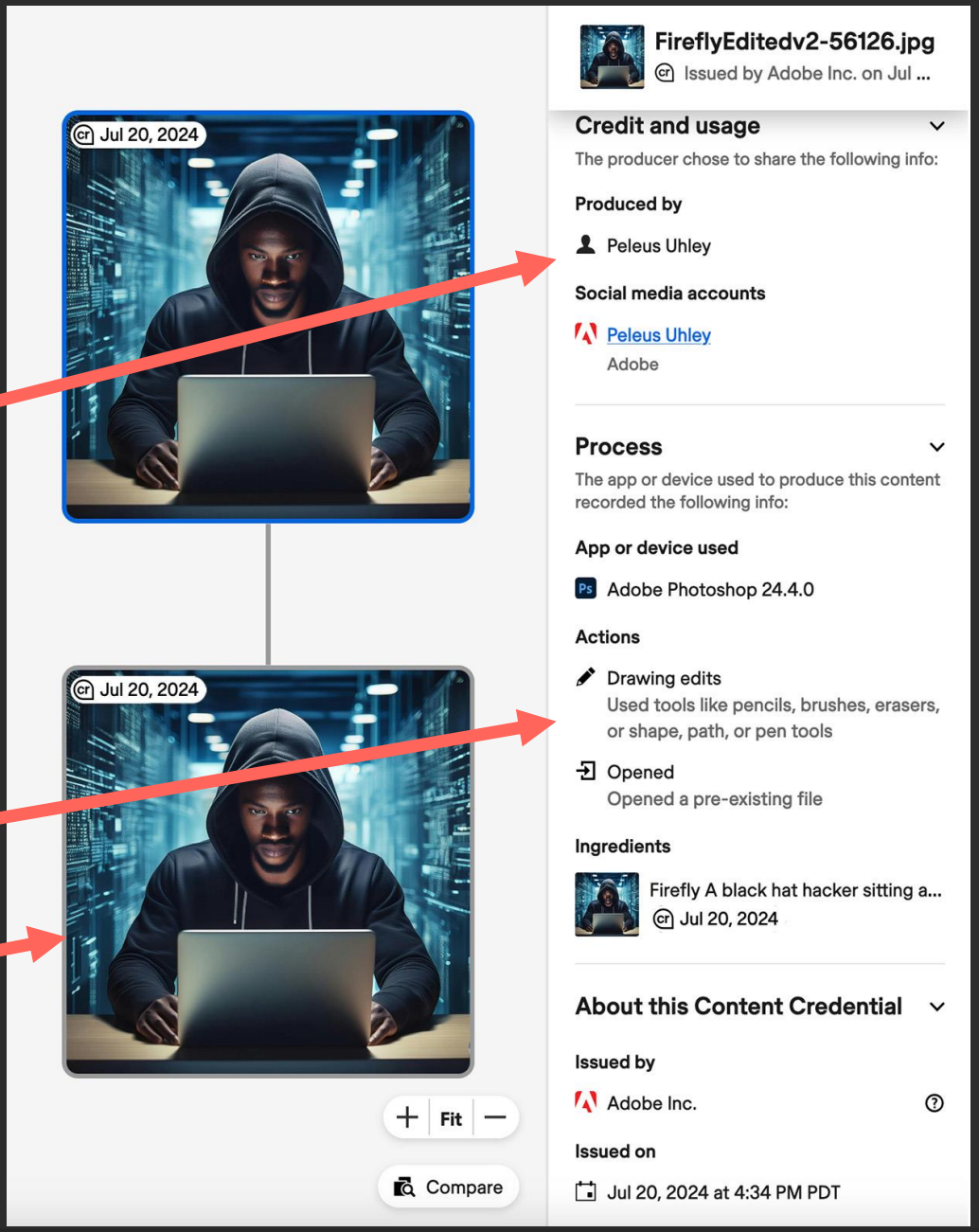
AI generated images of hackers created by Adobe Firefly





Before

After



Content credentials displayed in Microsoft PowerPoint

Click to add title

Content credentials provide confident information about the image's history and origin.

Created with Adobe Firefly
July 20, 2024 at 2:40 PM

2/3 3/4

Title area

Content credentials provide confident information about the image's history and origin.

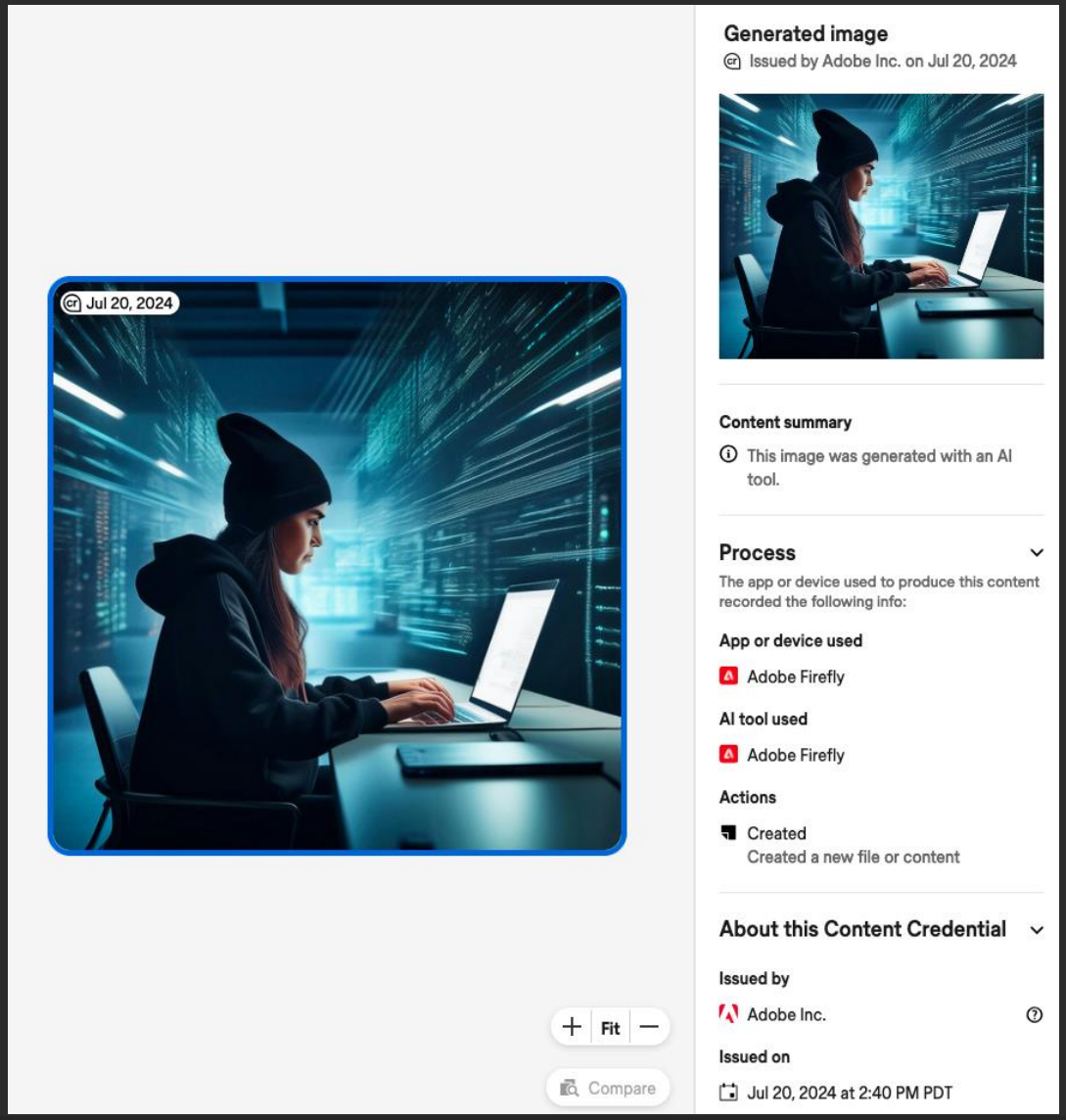
Created with Adobe_Photoshop/ 24.4.0 (build 20230411.r.433 c582fe9; mac) cai-helper/0.5.3 adobe_c2pa/0.2.1 c2pa-rs/0.16.0
July 20, 2024 at 4:34 PM

Content

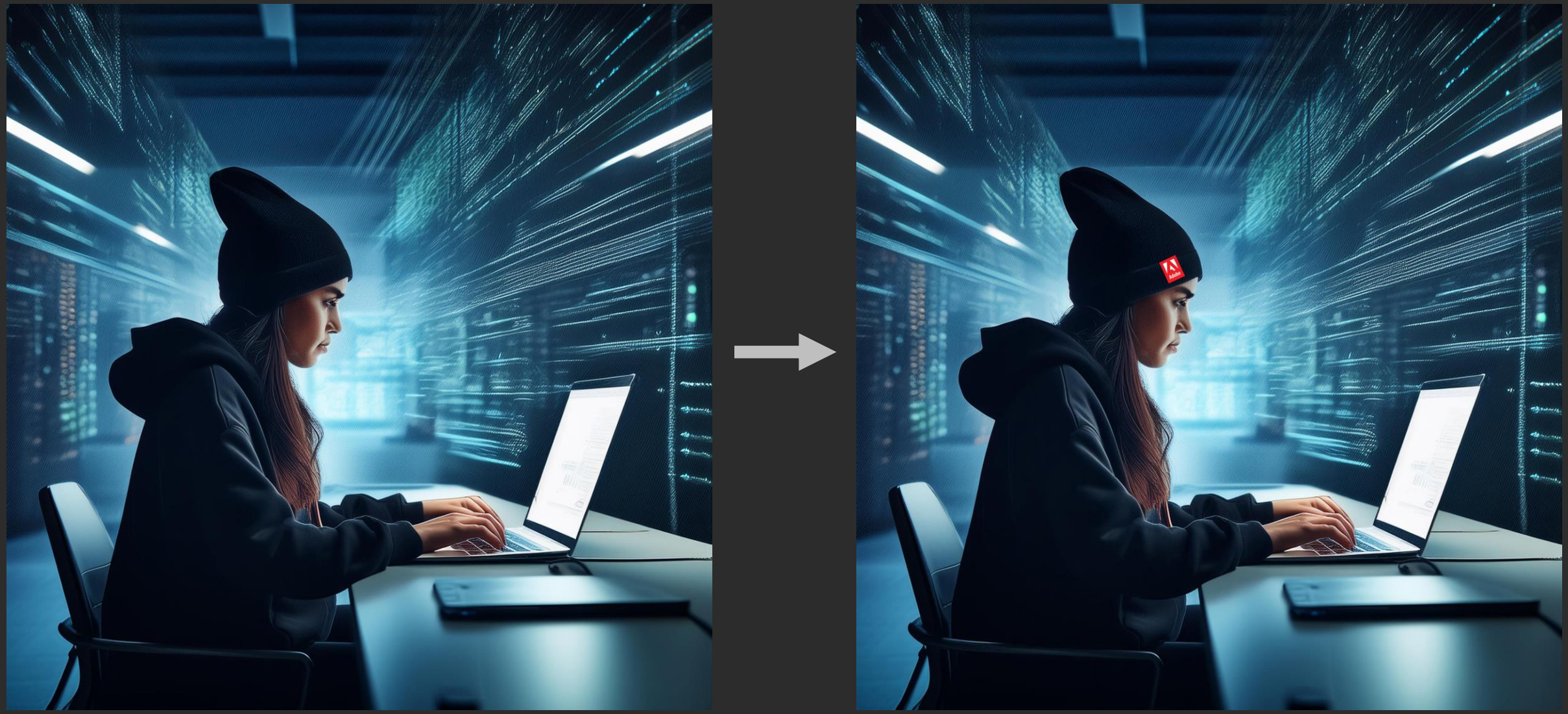
Adobe

Footer area

What happens when you edit without Content Credentials?

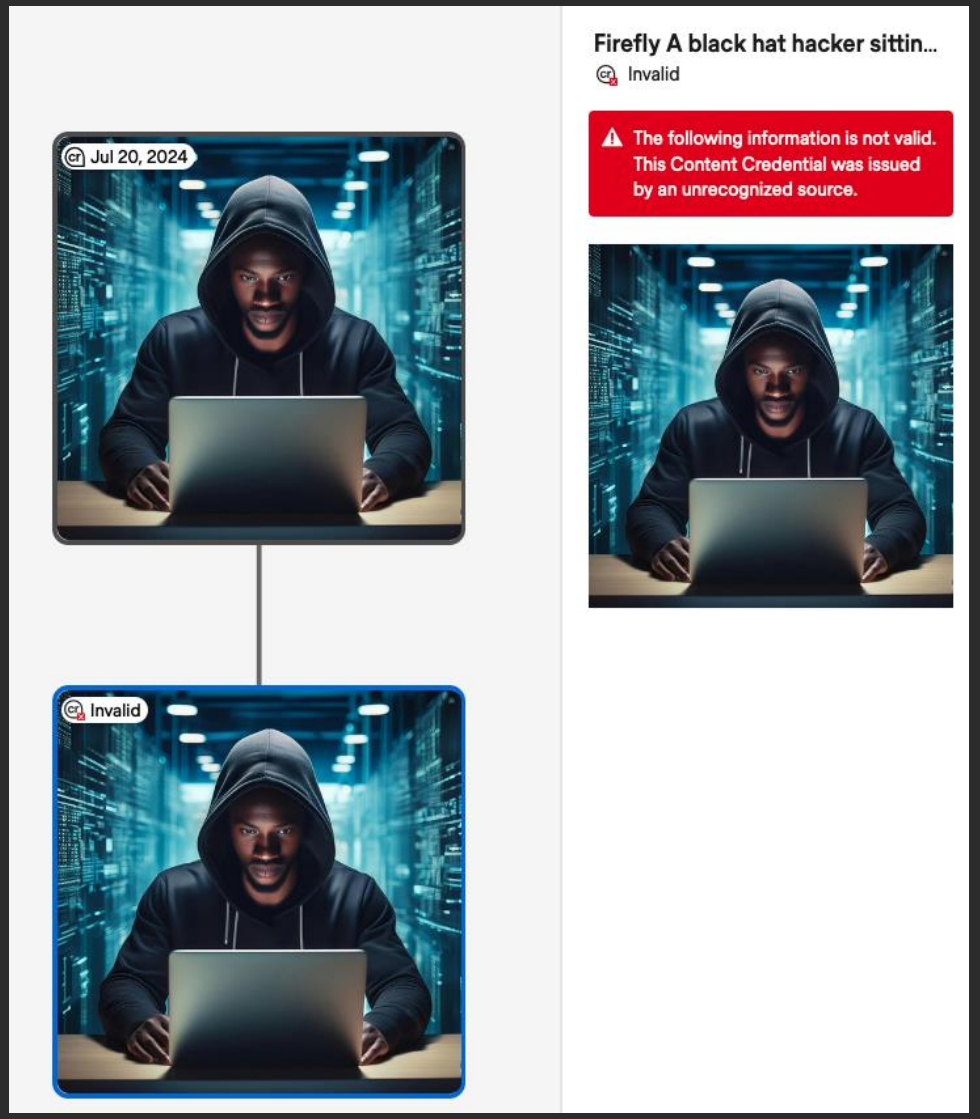


Modifying an image



Edits without using Content Credentials

Unrecognized or corrupted credentials



When?

Roadmap

- Rapidly advancing standard responding to community feedback
  - C2PA delivered version 1.0 of its technical standard for content provenance and authenticity in 2021. It has since updated that with version 1.1 and then 1.2 in 2022.
  - Versions 1.3 and 1.4 were released in 2023. Version 2.0 was released in January 2024
  - Version 2.1 will be released before the end of 2024. Currently a public draft on the C2PA website.
- The specification is outpacing the implementations
- Additional 2.0 guidance for the Security Considerations document and the c2pa-attacks tool

FAQs

Frequently asked questions

- Can’t I just delete the C2PA section of bits from the content? Yes.
- Will this be mandatory for all photos/videos/etc.? No. There are potential harms for minors, etc.
- Does this allow for anonymity and anonymous journalism? Yes *
- When will there be wide deployment? It is getting there but there are still parts in motion.
- Is this a watermark or fingerprint? No. C2PA is compatible with watermarking and fingerprinting.
- Is it a copyright? No. You can make the assertion, but C2PA isn't a copyright solution.
- Have there been bugs? Yes. We are rapidly iterating the spec in response to what we learned.
- Is there a bug bounty? Yes. Use the promo code "AdobeLoveBugBounty24" for a 35% bonus!

#1 FAQ: Will this all work?

- This model is similar to how you identify whether your desktop installers came from Microsoft.
- The goal is not to stop criminals from trying to commit crimes using deepfakes or perform deepfake detection
- The goal is to give media consumers methods to identify authentic content from the producers that they trust.
- The goal is to give media consumers methods to inspect the provenance of their content to understand what they are seeing and hearing.

Additional Resources

Opportunities to contribute

- Content Authenticity Initiative Bug Bounty
  - Funded and operated by Adobe
  - Includes open-source projects - https://github.com/contentauth/ with bug bounties
- Check to see if your company is already a member of C2PA, CAI, Project Origin, etc.
- We take feedback from public commentary
  - Adam Shostack reviewed the threat model, and his feedback was incorporated in the 1.4 version
  - File GitHub issues at: https://github.com/c2pa-org/

Resources

- C2PA specification: https://c2pa.org/specifications/specifications/2.0/specs/C2PA_Specification.html
- C2PA Security Guidance: https://c2pa.org/specifications/specifications/1.4/security/Security_Considerations.html
- C2PA Harms Modeling: https://c2pa.org/specifications/specifications/1.4/security/Harms_Modelling.html
- C2PA GitHub: https://github.com/c2pa-org/
- Creative Assertions Working Group: https://creator-assertions.github.io/
- Verify sites: https://contentcredentials.org/verify and https://contentintegrity.microsoft.com/
- c2pa-attacks testing tool: https://github.com/contentauth/c2pa-attacks

Questions

- https://twitter.com/peleusuhley
- https://bsky.app/profile/peleusuhley.bsky.social
- Email: puhley { @ } adobe.com
- CAI on Discord: https://discord.com/invite/CAI

- Test your skills: Real or Photoshop?

https://landing.adobe.com/en/na/products/creative-cloud/69308-real-or-photoshop/index.html