Authenticity
Authenticity from the attacker’s perspective
About Me

Dr. Neal Krawetz
Hacker Factor
Online Services

FotoForensics

hintfo

RootAbout

A Variety of Honeypots
Fraud!

News Outlets
Insurance

Fraud!

Banking

Shipping / Delivery

News Outlets

Different industries

Covid

(“I’m sick today”)

Injury

Propaganda

Ransom / Proof-of-Life

Don’t forget:

Political, Medical,
Scientific Research,
Legal Evidence,
Reputation, KYC,
Passports, Licenses,
Catfish, Celebrities,
Memes, UFOs, ...
Fraud!

Different methods

Deep Fakes

Insurance

Advanced Fakes

Banking

Simple Fakes

Shipping / Delivery

Dep Fakes

News Outlets

Staged

Altered

Covid ("I'm sick today")

Misrepresented

Injury

Ransom / Proof-of-Life

Propaganda

Don't forget:
- Political, Medical,
- Scientific Research,
- Legal Evidence,
- Reputation, KYC,
- Passports, Licenses,
- Catfish, Celebrities,
- Memes, UFOs, ...
Three-Legged Foundation
Provenance

- **Content**
- **Metadata**
  - EXIF, XMP, IPTC
  - MakerNotes
  - Digests, Checksums
Provenance

- Content
- Metadata
  - EXIF, XMP, IPTC
  - MakerNotes
  - Digests, Checksums

Pirates say “ARRR!”
- Alter
- Remove
- Replace
- Re-encode
Currently, C2PA works primarily on images and video, though members say that they are working on ways to handle text-based content. I get into some of the other shortcomings of the protocol in the piece, but what's really important to understand is that even when the use of AI is disclosed, it might not stem the harm of machine-generated misinformation. Social media platforms will still need to decide whether to keep that information on their sites, and users will have to decide for themselves whether to trust and share the content.

It’s a bit reminiscent of initiatives by tech platforms over the past several years to label misinformation. Facebook labeled over 180 million posts as misinformation ahead of the 2020 election, and clearly there were still considerable issues. And though C2PA does not intend to assign indicators of accuracy to the posts, it’s clear that just providing more information about content can’t necessarily save us from ourselves.
Watermarks

- Watermarks
  - Invisible
  - Visible
Watermarks

- Watermarks
  - Invisible
    - DigiMarc
    - Stable Diffusion
Watermarks

- Watermarks
  - Invisible
  - Visible

- Must disclose method
  - Easy to erase
Watermarks

- Watermarks
  - Invisible
  - Visible

- Must disclose method
  - Easy to erase
  - Easy to add
    - False attribution
Meta’s AI Watermarking Plan Is Flimsy, at Best

Watermarks are too easy to remove to offer any protection against disinformation

BY DAVID EVAN HARRIS, LAWRENCE NORDEN | 04 MAR 2024 | 6 MIN READ |
Fingerprinting

- Ballistics / Profiling
- Complex Structures
- File Encoding Options
- Camera Artifacts
- Similar Image Search
Fingerprinting

- Ballistics / Profiling
  - Complex Structures
  - File Encoding Options
  - Camera Artifacts
- Similar Image Search

AvtoVAZ Method (circa 2013)
1. Alter a photo
2. Replicate fingerprints
End result? Forgery that looks original.
Falsified Photos: Fooling Adobe’s Cryptographically-Signed Metadata

by: Adam Zeloof

November 30, 2023

Based on Trust

<table>
<thead>
<tr>
<th>Traditional Media Analysis:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
</tr>
</tbody>
</table>

**Today’s Forensic Examiners:**

**Trust but Verify**

Tools, techniques, and methods that check for consistencies. *Inconsistencies* are indicators of alterations or tampering.
### Traditional Media Analysis:

<table>
<thead>
<tr>
<th>Content</th>
<th>Assume unaltered or acceptable alterations, not misrepresented.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metadata</strong></td>
<td><strong>Trust</strong> metadata accurately reflects the content.</td>
</tr>
<tr>
<td></td>
<td>Relies on the <strong>honesty</strong> of the person inserting the metadata.</td>
</tr>
</tbody>
</table>

#### C2PA adds:

<table>
<thead>
<tr>
<th><strong>Metadata</strong></th>
<th><strong>C2PA</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera Make/Model</td>
<td>Manifest</td>
</tr>
<tr>
<td>Software Version</td>
<td>C2PA Metadata</td>
</tr>
<tr>
<td>Date / Time</td>
<td>Assertions</td>
</tr>
<tr>
<td>Location</td>
<td>Signed Claims</td>
</tr>
<tr>
<td>Photographer</td>
<td>Certificates</td>
</tr>
<tr>
<td>Copyright</td>
<td>Signatures</td>
</tr>
<tr>
<td>Description</td>
<td>Notary Timestamp</td>
</tr>
</tbody>
</table>
## Based on Trust

<table>
<thead>
<tr>
<th>Traditional Media Analysis:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content</strong></td>
<td>Assume unaltered or acceptable alterations, not misrepresented.</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td>Trust metadata accurately reflects the content.</td>
</tr>
<tr>
<td></td>
<td>Relies on the honesty of the person inserting the metadata.</td>
</tr>
<tr>
<td><strong>C2PA adds:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>C2PA Metadata</strong></td>
<td>Trust that it accurately reflects the content.</td>
</tr>
<tr>
<td><strong>Certificate</strong></td>
<td>Trust certificate is issued to authorized source.</td>
</tr>
<tr>
<td><strong>Signer</strong></td>
<td>Trust that signers validated the metadata and content; <strong>not required</strong>.</td>
</tr>
<tr>
<td></td>
<td>Trust new signers didn’t alter previous claims.</td>
</tr>
<tr>
<td><strong>Validation</strong></td>
<td>Trust tools to perform proper validation.</td>
</tr>
<tr>
<td></td>
<td>Trust signature covers entire file; <strong>not required</strong>.</td>
</tr>
<tr>
<td></td>
<td>Trust “tamper evident” detects tampering.</td>
</tr>
<tr>
<td><strong>Peer Pressure</strong></td>
<td>Trust that thousands of reviewers actually reviewed it.</td>
</tr>
</tbody>
</table>
Live Demo!

How to create an authenticated C2PA forgery in under one minute!
Traditional Media Forensics

I can prove the file is altered!
With C2PA

I can prove the file is altered!
& signature is untrusted!
& “tamper evident” crypto failed!
& the 100s of companies that claim C2PA works are wrong!

Regular users will blindly trust C2PA!
Other Solutions? (besides C2PA)

<table>
<thead>
<tr>
<th>Solution Approach</th>
<th>Attack Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vendor Dependent</td>
<td>DoS: Knock the vendor offline or discredit</td>
</tr>
<tr>
<td>Computationally Bound (e.g., blockchain)</td>
<td>Flood with forgeries, scaling issues, inherent delays for timely validation</td>
</tr>
<tr>
<td>Time-based Solution</td>
<td>Backdate or postdate</td>
</tr>
<tr>
<td>Registration-based Solution</td>
<td>Register first, or contest prior registration</td>
</tr>
<tr>
<td>Hardware-integrated</td>
<td>Replace the hardware, inject into workflow</td>
</tr>
<tr>
<td>Cost restrictions, Entrance fees</td>
<td>Fraud is a $Billion industry!</td>
</tr>
</tbody>
</table>
Takeaways

- Authentication, Provenance, Validation, Vetting
  - Hard problems
  - No “easy button” or simple solution

- Attackers aren’t stupid
  - “Trust” and “Honesty” are easy targets
  - Vulnerabilities will be exploited
Authenticity from the attacker’s perspective

Dr. Neal Krawetz
Hacker Factor