

# Authenticity

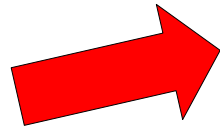


# Authenticity from the attacker's perspective



# About Me

Dr. Neal Krawetz  
Hacker Factor



# Online Services



RootAbout



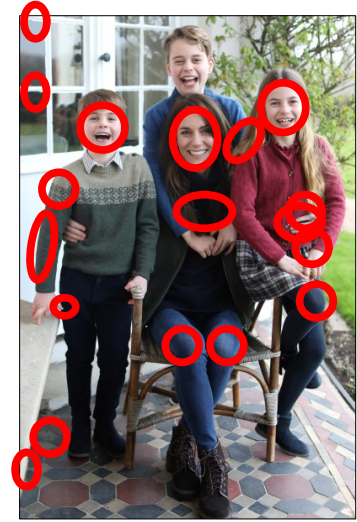
A Variety of  
Honeypots

Fraud!



News Outlets

Fraud!



News Outlets

# Insurance



# Fraud!

# Different industries



Banking

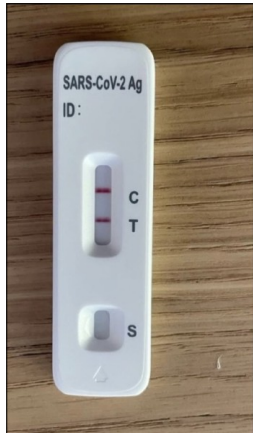


Shipping / Delivery



News Outlets

# Covid ("I'm sick today")



# Injury



Ransom / Proof-of-Life



Propaganda

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



# Insurance



Deep Fakes



Simple Fakes

# Fraud!



Advanced Fakes

## Shipping / Delivery

# Different methods



Altered

## News Outlets



Misrepresented

Covid ("I'm sick today")

Injury



Ransom / Proof-of-Life



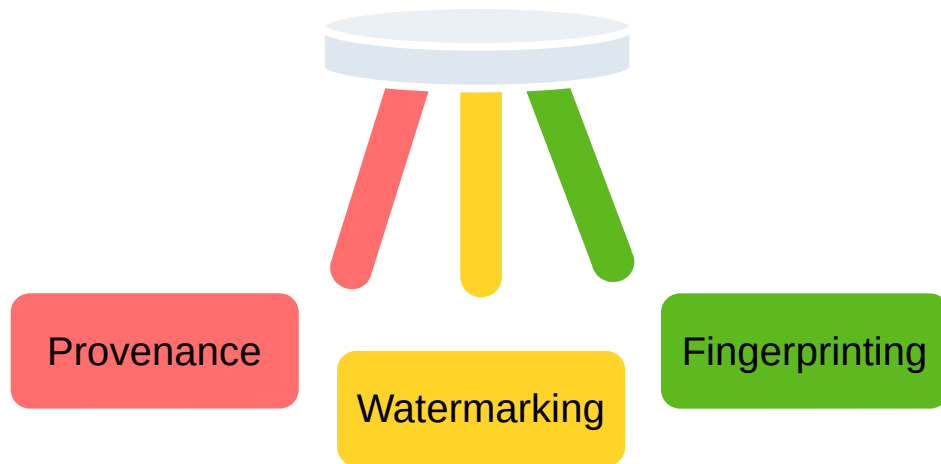
Staged

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



Pirates say “ARRR!”

- Alter
- Remove
- Replace
- Re-encode



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



# Provenance

## MIT Technology Review

### POLICY

## The race to find a better way to label AI

An internet protocol called C2PA uses cryptography to label images, video, and audio

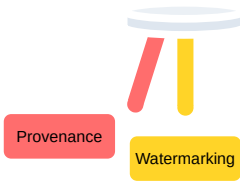
By Tate Ryan-Mosley

July 31, 2023

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.

<https://www.technologyreview.com/2023/07/31/1076965/the-race-to-find-a-better-way-to-label-ai/>

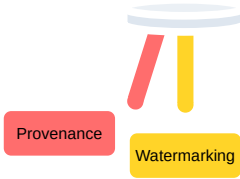


# Watermarks



- Watermarks
  - Invisible
  - Visible



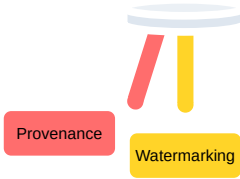


# Watermarks



- Watermarks
  - Invisible
    - DigiMarc
    - Stable Diffusion



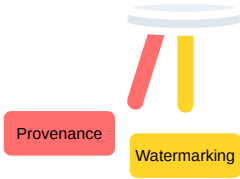


# Watermarks



- Watermarks
  - Invisible
  - Visible
- Must disclose method
  - Easy to erase



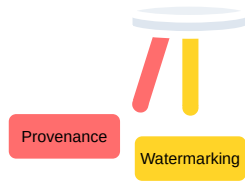


# Watermarks



- Watermarks
  - Invisible
  - Visible
- Must disclose method
  - Easy to erase
  - Easy to add
    - False attribution





# Watermarks

**IEEE Spectrum** FOR THE TECHNOLOGY INSIDER

GUEST ARTICLE | **ARTIFICIAL INTELLIGENCE**

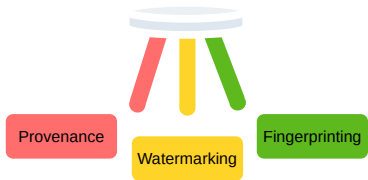
## **Meta's AI Watermarking Plan Is Flimsy, at Best** > Watermarks are too easy to remove to offer any protection against disinformation

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

<https://spectrum.ieee.org/meta-ai-watermarks>







# Fingerprinting

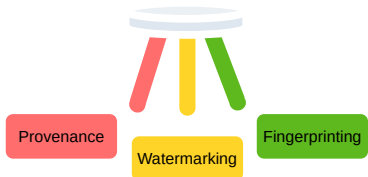
- Ballistics / Profiling
  - Complex Structures
  - File Encoding Options
  - Camera Artifacts
- Similar Image Search



## AvtoVAZ Method (circa 2013)

1. Alter a photo
2. Replicate fingerprints

**End result?** Forgery that looks original.



# Fingerprinting



# HACKADAY

HOME

BLOG

HACKADAY.IO

TINDIE

HACKADAY PRIZE


SUBMIT

ABOUT

March 10, 2024

## FALSIFIED PHOTOS: FOOING ADOBE'S CRYPTOGRAPHICALLY- SIGNED METADATA

by: [Adam Zeloof](#)

 [34 Comments](#)



November 30, 2023

<https://hackaday.com/2023/11/30/falsified-photos-fooling-adobes-cryptographically-signed-metadata/>



# Based on Trust

Traditional Media Analysis:	
<b>Content</b>	<b>Assume</b> unaltered or acceptable alterations, not misrepresented.
<b>Metadata</b>	<b>Trust</b> metadata accurately reflects the content. Relies on the <b>honesty</b> of the person inserting the metadata.

Today's Forensic Examiners:

**Trust but Verify**



Tools, techniques, and methods that check for consistencies.  
*Inconsistencies* are indicators of alterations or tampering.

# Based on Trust

## Traditional Media Analysis:

<b>Content</b>	<b>Assume</b> unaltered or acceptable alterations, not misrepresented.
<b>Metadata</b>	<b>Trust</b> metadata accurately reflects the content. Relies on the <b>honesty</b> of the person inserting the metadata.

## C2PA adds:



### Metadata

Camera Make/Model  
Software Version  
Date / Time  
Location  
Photographer  
Copyright  
Description

### C2PA

Manifest  
C2PA Metadata  
Assertions  
Signed Claims  
Certificates  
Signatures  
Notary Timestamp

# Based on Trust

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

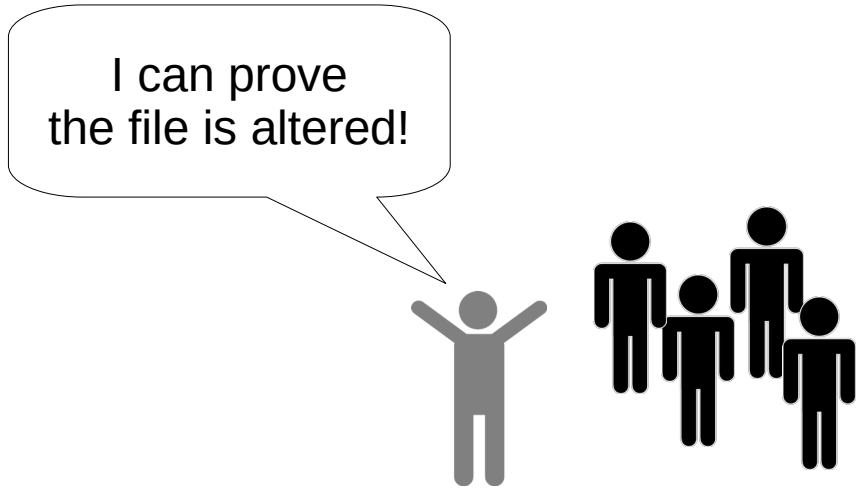


# Live Demo!



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

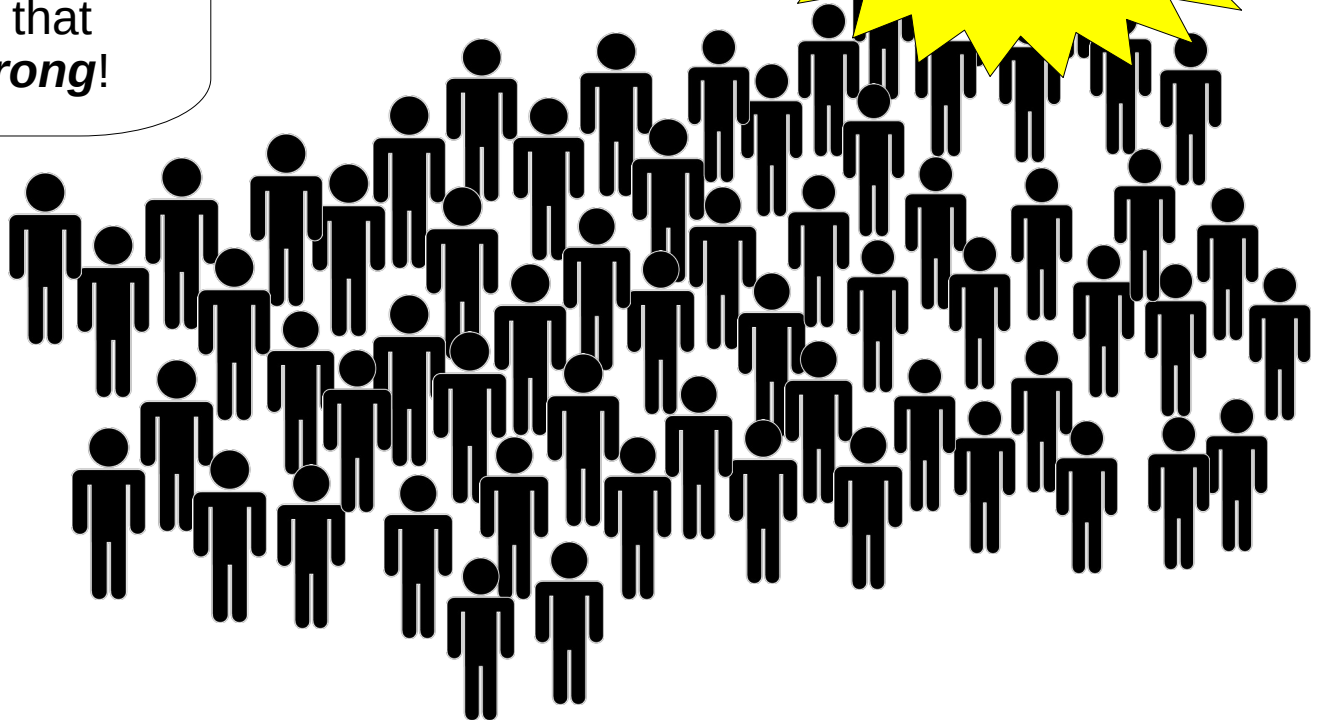
# Traditional Media Forensics



# 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)

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



# 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



Provenance

Watermarking

Fingerprinting

# Authenticity from the attacker's perspective

Dr. Neal Krawetz  
Hacker Factor

