Harms Modelling in the C2PA
Coalition for Content Provenance and Authenticity

Jacob Castellanos Rivadeneira
Technology Threats and Opportunities at WITNESS
jacobo@witness.org
@JacoboCas
Agenda

1. Background on WITNESS and our work on Provenance and Authenticity Infrastructure

2. Overview of the C2PA (Coalition for Content Provenance and Authenticity)

3. Harms Modelling in the C2PA
Provenance and Authenticity (P&A) Infrastructure

Provenance and Authenticity Infrastructure refers to the tools, services or frameworks that facilitates capturing, processing and presenting information about the source and history of digital assets in a way that is verifiable and tamper-evident.
WITNESS work on **Authenticity Infrastructure**: identifying values and importance, highlighting trade-offs

Other resources:

- Synthetic Media Lab
From niche to systemic P&A Infrastructure
As we move towards systemic use…

How do we prevent, avert and mitigate harm?

How do we enhance freedom of expression and trust?
The Coalition for Content Provenance and Authenticity (C2PA) is an initiative that addresses the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of digital assets.
The specifications intend to offer a secure way to establish the provenance of digital assets across platforms...

Content with provenance provides **indicators of authenticity** so that consumers can have awareness of who has altered content and what has been changed.
2 objectives within the C2PA
Prevent, avert, mitigate harms (Bolstering human rights framework)
Promote critical usages

Steering Committee

Technical Working Group

Harms reporting / continuous assessment
Resourcing / Supporting a diverse C2PA ecosystem
Guiding Principles
Threats and Harms Taskforce
Harms Modelling

Strategies for averting, mitigating harms
The Guiding Principles of the C2PA

- Privacy
- Global Audience / Accessibility
- Simplicity and Cost Burden
- Misuse

- C2PA specifications MUST respect the common privacy concerns of each of the target users named earlier.
  o C2PA specifications MUST allow content creators, editors, and publishers to remove sensitive information before sharing with others. Subsequent participants must be made aware of such removal.
  o C2PA specifications MUST NOT require identity of the person or organization making any assertion or claim about an asset to be documented. The specifications MAY allow that information to be represented, provided that representation is optional.

- C2PA specifications MUST take into consideration the needs of interested users throughout the world.
  o C2PA-aware tools SHOULD be accessible to users with limited or high-cost access to Internet services.

- C2PA specifications MUST be reviewed with a critical eye toward potential abuse and misuse of the framework.

...
Harms Modelling
The Harms, Misuse, and Abuse Assessment is an ongoing process that accompanies the design, development, implementation and use stages of the C2PA standard, and the process includes a multi-disciplinary and diverse range of stakeholders.
Assessment Methodology

- **Internal consultations & discussions**
- **TWG**
- **External consultations**
- **Threats & Harms TF**

Focus on global stakeholders, with different technical, lived, practical or professional experiences + most likely affected and marginalized from these processes.
## Overview of results

<table>
<thead>
<tr>
<th>Category</th>
<th>Type of Harm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denial of consequential services</td>
<td>Opportunity loss (5)</td>
</tr>
<tr>
<td></td>
<td>Economics loss (4)</td>
</tr>
<tr>
<td>Infringement on human rights</td>
<td>Dignity loss (1)</td>
</tr>
<tr>
<td></td>
<td>Liberty loss, discrimination and lack of due diligence (6)</td>
</tr>
<tr>
<td></td>
<td>Privacy loss (5)</td>
</tr>
<tr>
<td></td>
<td>Constraints on freedom of expression (2)</td>
</tr>
<tr>
<td></td>
<td>Freedom of associations, assembly and movement (2)</td>
</tr>
<tr>
<td></td>
<td>Environmental impact (1)</td>
</tr>
<tr>
<td>Erosion of social and democratic structures</td>
<td>Manipulation (6)</td>
</tr>
<tr>
<td></td>
<td>Over-reliance on technical systems (1)</td>
</tr>
<tr>
<td></td>
<td>Social detriment (2)</td>
</tr>
<tr>
<td>Risk of injury</td>
<td>Emotional or psychological distress or physical harm (1)</td>
</tr>
</tbody>
</table>
Reduction in options for anonymity and pseudonymity

Privacy loss
Human rights activist inadvertently includes location in media assertion and is subsequently targeted (c.f. existing precedents of inadvertent release of metadata, most famously John McAfee or recurring cases in human rights).

Attacks on journalistic freedom and independence

Opportunity loss
An abuse of the C2PA system to enforce journalistic identity in laws in a jurisdiction or demand additional information on media posted on social media leads to a reduction of media diversity and suppression of speech.

Requiring participation in the use of technology or surveillance to take part in society

Freedom of association, assembly and movement.
For example, algorithmic ranking: content creators forced to game algorithms with particular keywords, metadata to achieve visibility/to be ranked higher in a feed.
Potential harm deriving from soft-binding and the use of manifest datastores

Rendition of an image that does not contain C2PA manifests.

Soft-binding extracted (e.g. perceptual hash)

Interactive search through a provenance datastore

Database includes manifests and thumbnails

Thumbnail match shown to consumer for validation.
Trust model threats and harms

**Claim generator** (Implementation)
- **Signer**: Trusts issuer to identify signer
- **Identity issuer**: Trusts signer to secure its credentials
- **Consumer**: Trusts assertions are made by the signer
- **Validator**: Trusts validator to validate and correctly identify signer

**Spoofing**: Credential is misleading, e.g., BBC_LA

**Intended use**: If you are unknown entity (e.g., La Sandia), signer is irrelevant because there is no trust in signer (no value for civic media etc.)

**Unintended use**: Signed manifest is taken as trust, not trust signal.

**Tampering**: either an update manifest, legacy media or security breach

**Holds trust lists of CAs**

**Cost of x.509 certificates raises accessibility concerns**

**Credentials compromised, e.g., stolen or revoked**

**Technical difficulty to obtain a certificate raises accessibility concerns**

**There is no guidance for maintaining trust lists**

**Unclear UX misrepresents trust signals**

**Credential is irrelevant to consumer**

**Intended use**: If you are unknown entity (e.g., La Sandia), signer is irrelevant because there is no trust in signer (no value for civic media etc.)

**Unclear UX misrepresents trust signals**

**Unintended use**: signed manifest is taken as trust, not trust signal.
<table>
<thead>
<tr>
<th>Category</th>
<th>Type of Harm</th>
<th>Potential Harm / Misuse / Abuse</th>
<th>Contextual Example / Evidence</th>
<th>Existing and Potential Mitigations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>discrimination</td>
<td>Limited language versioning on CIPA-enabled tools, despite their focus on low-cost and global accessibility, leads to more limited access for marginal markets.</td>
<td>CPA-enabled tools are likely to leave out languages with marginal markets. A parallel example is that of the continued use in Myanmar of Zawgyi as the dominant typeface used to encode Burmese language characters rather than Unicode, the international text encoding standard, resulting in technical challenges for many companies that provide mobile apps and services.</td>
<td>Specifications: Manifest localization specifications to be added beyond 1.0. No other limitations have been identified on language versioning of CIPA implementations. Non-technical and multilateral harms response actions: The harms, misuse and abuse assessment should guide potential multilateral cooperation for the promotion of a diverse CIPA ecosystem, including language-inclusive implementations.</td>
</tr>
<tr>
<td></td>
<td>Digital divide/technological discrimination (1)</td>
<td>Individuals and communities using older devices or operating systems as creators/consumers or using access to the internet via Free Basics or equivalent “affordable access” approaches that limit the websites and services an actor can access.</td>
<td>For example, existing experiences with gated/limited access to particular websites and tools via Free Basics program for “affordable access” from mobile operators in emerging markets. See also example above on Educational discrimination and limited language versioning.</td>
<td>Specifications: Specifications do not preclude CIPA implementations in older devices and operating systems. Specifications are open, global and opt in. The specifications use open standards for which there are existing libraries in various programming languages across a range of devices and operating systems/environments. To facilitate access for individuals or communities who do not, or cannot, have access to x.509 certificates, the specifications allow for self-signing certificates. Accompanying documentation and guidance: Minimum viable implementations guidance to be developed as a fallback for older devices and operating systems. Guidance for implementers includes recommendations on the use of a private credential store (also known as the &quot;address book&quot;).</td>
</tr>
<tr>
<td></td>
<td>Digital divide/technological discrimination (2)</td>
<td>Individuals and communities without ability to access or use tools for compliance with system usage are excluded.</td>
<td>Financial costs involved in signing up to use different CIPA-enabled tools and software may exclude marginalized individuals and communities who cannot afford the cost. For example, exclusion of content creators without compliant x.509 certificates. Lack of literacy and access to education about the tool may also limit usage among marginalized populations.</td>
<td>Non-technical and multilateral harms response actions: The harms, misuse and abuse assessment should guide potential multilateral cooperation for the promotion of a diverse CIPA ecosystem and encourage the development of simple products to meet claim generation and validation requirements in diverse environments. An ongoing harm assessment should inform the continuous development of the specifications to address issues that limit CIPA implementations in older devices and operating systems.</td>
</tr>
<tr>
<td>Opportunity loss</td>
<td></td>
<td>Journalistic Freedom and Independence</td>
<td>An abuse of the CIPA system to enforce journalistic identity laws in a jurisdiction or demand additional information on media posted on social media leads to a reduction of media diversity and suppression of speech. Misuse of manifest repositories to track content or enforce restrictive laws on freedom of expression and do so with lack of effective remedy and/or exploitation of manifest repositories to track content, and curtail freedom of expression (e.g. political speech).</td>
<td>Specifications: Specifications are open, global and opt in. If they are used, the CIPA provides features that can be used to protect confidentiality of personal information while still establishing the provenance of an asset, including anonymous and pseudonymous signing, redaction as an authorized action, use of update manifests with redacted information, and the use of VSC credentials. No sensitive information is required in CIPA workflows. Accompanying documentation and guidance: User experience guidance provides recommendations to prevent inadvertent disclosure of information. Guidance for implementers highlights trust and privacy considerations, including on the use of manifest repositories: We recommend that claim generators that add soft binding assertions to an asset’s manifest do so as an opt-in addition and not make it mandatory. Guidance also recommends that content creators be informed of the trade-offs involved in using manifest repositories that allow for asset link-up with soft bindings; that is, on the one hand, identifying manifests that have become &quot;decoupled&quot; from their associated assets, while on the other hand, privacy risks that may result from a soft binding link-up to an earlier manifest with, for example, redacted information. Non-technical and multilateral harms response actions: The harms, misuse and abuse assessment should inform the CIPA to proactively engage and lobby</td>
</tr>
</tbody>
</table>

See overlap with Journalistic Pluraity and Diversity
Outputs of Harms Modelling

Harm, Abuse, Misuse Assessment

Specifications
- UX Guidance
- Guidance for implementers
- Explainer

Accompanying documents

Non-technical and multilateral harms response actions
- Ongoing Assessment
- Governance
- Resourcing / Supporting diverse ecosystem