



# **Coalition for Content Provenance and Authenticity**

## C2PA Harms Modelling

1.4, 2024-08-22: Release

# Table of Contents

- 1. Harms, Misuse, and Abuse Framework ..... 2
- 2. Methodology ..... 4
  - 2.1. Phase I: Purposes, Use-cases, Users and Stakeholders..... 4
  - 2.2. Phase II: Harm Taxonomy and Assessment ..... 4
  - 2.3. Phase III: Due Diligence Actions ..... 5
- 3. Harms, Misuse, and Abuse Initial Assessment ..... 6
  - 3.1. Phase I: Purposes, Use-cases, Users and Stakeholders..... 6
  - 3.2. Phase II: Harm Taxonomy and Assessment ..... 6
  - 3.3. Phase III: Due Diligence Actions ..... 7
- 4. Public Review and Feedback ..... 9
- 5. Due Diligence Actions ..... 10
- 6. Harms considerations for C2PA stakeholders ..... 11
  - 6.1. General considerations for content creators..... 11
  - 6.2. General considerations for content consumers (consumers of Content Credentials)..... 12
  - 6.3. General considerations for civic, community, and independent media ..... 13
  - 6.4. General considerations for human rights defenders ..... 15



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

# Chapter 1. Harms, Misuse, and Abuse Framework

Harms modelling focuses on analysing how a socio-technical system might negatively impact users, stakeholders, broader society, or otherwise create or re-enforce structures of injustice, threats to human rights, or disproportionate risks to vulnerable groups globally. The process of harms modelling systematically requires combining knowledge about a system architecture and its user affordances, with historical and contextual evidence about the impact of similar existing systems on different social groups. This combined information frames the ability to anticipate harm.

Harms modelling considers the ramifications of a technological system both from the perspective of the technology developers as well as users and non-user stakeholders. In other words, harms modelling considers what kinds of harms may result from the configuration of a system as well as what kinds of harms may result from both its intended use and unintended use. It is necessary to combine all of these considerations to achieve a broader perspective on potential harms, particularly on those that may be unanticipated by system developers but highly evident to disproportionately impacted social groups. Following principles from justice-oriented technology development and design justice, it is essential to include wide-ranging and ongoing consultations with communities likely to be impacted by the specification, and to place emphasis on those who already face similar systemic harms.

In designing our harms modelling approach, the Taskforce drew inspiration from different approaches to technology impact assessment, including the fields of value-driven design, human rights due diligence, security-focused threat modelling frameworks, and harms modelling methodologies. After a review of potential methodologies, it was determined that adapted versions of [Microsoft's Harms Modelling Framework](#) and [BSR's Human Rights Due Diligence Assessment](#) would be used to guide the harms modelling process for C2PA. The Taskforce also collaborated with colleagues conducting parallel exercises in threat modelling exercises as part of the Technical Working Group and engaged with the User Experience research and Decoupled Taskforces within C2PA.

Some of the modifications made to existing frameworks for technology impact assessment are listed below:

## **Human rights-focused harm taxonomy**

The Taskforce sought to ensure that more well-established human rights, privacy and security concerns were analysed as elements of broader forms of harm around social inequality and discrimination, and in relation to issues potentially affecting the particular users and stakeholders of the C2PA (such as media entities, citizen journalists, and human rights defenders). For this reason, the harm taxonomy particular to the Microsoft Harms Modelling Framework was modified to reflect these issue intersections, stakeholders, and users. The reader will note that some harm taxonomy categories are broader than others. This reflects the fact that there is significant overlap between categories and using both broad and narrow categories helped to consider a range of potential harms, misuses and abuses.

## **Temporality**

It is important to analyze harms and impacts not as a static snapshot in time but as an ongoing process with particular considerations for every stage of technological design, development and use (and potentially non-use). This is reflected in the following scenarios of analysis: 1) Initial Adoption; 2) Wide Adoption and 3) Ongoing Maintenance. These scenarios are explained further in the Harms, Misuse, and Abuse Initial Assessment section.

### **Assigning values for severity, scale, likelihood, frequency and impact**

The Taskforce conducted an internal process to understand severity, scale, likelihood, frequency and impact of potential harms. This was done in consultation with issue experts within the C2PA and based on C2PA member WITNESS's work and consultation globally on trade-offs and risks within authenticity and provenance infrastructure ([see Ticks or It Didn't Happen: Confronting Key Dilemmas in Building Authenticity Infrastructure for Multimedia](#)).

Further consultation followed with outside groups, particularly with communities with lived, practical and expert knowledge, and with those who may be disproportionately impacted by potential harms and that are often most excluded from design.

This analysis will be ongoing, considering that the degree of severity, likelihood, and impact will likely change and become more evident after the specifications are implemented into products and deployed.

### **Considering accountability**

Acknowledging that ethical analyses and threat modelling processes are sometimes done behind closed doors, it is important to emphasize that the harm assessment will be continuously inclusive and it will inform future specifications development, the governance of the coalition, potential parallel compliance mechanisms, and cooperation and resourcing for a diverse C2PA ecosystem.

# Chapter 2. Methodology

There are three phases to our methodology. These phases do not reflect a chronological order, they frame specific processes that will need to be continuously iterated as more actors join the discussion and analysis, both before and after the publication of version 1.0.

## 2.1. Phase I: Purposes, Use-cases, Users and Stakeholders

Phase I includes defining the purposes of the technology, its use-cases and stakeholders as it pertains to the harms, misuse and abuse assessment. As with other parts of the C2PA standards development, the Taskforce began with the Purposes/Use-Cases/Users/Stakeholders from two initiatives, the [Content Authenticity Initiative](#) and [Project Origin](#) and expanded to other potential scenarios.

Some of the questions to be addressed were:

### Purposes

What problem will be solved? For who? What new capability will be possible? For who?

### Use-cases

What will the C2PA standard be used for? What context will the C2PA standard likely be used in?

### Users/Actors

Who will directly interact with the C2PA standard?

### Stakeholders

Who will be impacted by the use of the C2PA standard including non-users?

## 2.2. Phase II: Harm Taxonomy and Assessment

In Phase II, the Taskforce reviewed and adapted Microsoft's taxonomy of harm to better reflect the context and implications of the C2PA specifications. This process was intertwined with the actual assessment of the identified harms. The guiding questions in this phase were:

- How could people be harmed by the use of C2PA? What use-cases are most likely to cause harm? To whom?
- What use-cases are most likely to cause harm? To whom?
- How could a misuse or abuse of C2PA lead to harm? Who would be affected?
- What contextual evidence from either an existing technology or societal phenomenon either provides direct evidence of this harm or harm in a related context?
- What is the severity, scale, frequency, likelihood, and disproportionate impact on vulnerable groups of a particular potential harm?

## 2.3. Phase III: Due Diligence Actions

Phase III was aimed at mitigating potential abuse and misuse, and offering considerations and guidelines for the protection of human rights and for the optimization of the benefits that prompted the development of the C2PA standard. The questions that guided us in this phase were:

- How could the C2PA specifications be designed to prevent harmful impacts?
- How could the C2PA specifications be built to protect human rights?
- What guidance, compliance requirements or technical steps can address these?

Answers to these questions are reflected in the due diligence strategy that affects the specifications and its accompanying documents, which includes guidance for implementers, guidance on user experience, security considerations, and an explainer aimed at the general public.

Due diligence recommendations resulting from the harm assessment should also inform the governance of the Coalition and guide potential multilateral cooperation for the promotion of a diverse C2PA ecosystem that pushes for the optimization of the benefits in terms of trust in media, user control and transparency that prompted the development of the C2PA specifications.

# Chapter 3. Harms, Misuse, and Abuse Initial Assessment

The harms, misuse, and abuse assessment is an ongoing process. The information presented in the Harms Modelling documentation should not be considered the end result of a comprehensive evaluation, but as a basis for ongoing discussions centred on impacted communities, and aimed at mitigating potential abuse and misuse and protecting human rights.

There are two critical aspects of the approach:

## Ongoing

The harms, misuse, and abuse assessment necessarily accompanies the design and development, as well as implementation and use-stages of the C2PA by continuously informing the specifications development process, the implementation and user-experience guides, sensitization efforts, the governance of the Coalition and potentially multilateral cooperation for the promotion of a diverse C2PA ecosystem that serves a broad range of global contexts.

## Multi-disciplinary and diverse

The harms, misuse, and abuse assessment is a collaborative effort that includes multi-disciplinary experts and a broad range of stakeholders with lived, practical and technical experience of the issues and from diverse geographical locations, cultural backgrounds and individual identities.

## 3.1. Phase I: Purposes, Use-cases, Users and Stakeholders

For more information on purposes and use-cases of the C2PA specifications, see the examples listed in the [Explainer](#). Note that this is not an exhaustive list, but an extension of use-cases that have come up in parallel organizations such as the [Content Authenticity Initiative](#) and [Project Origin](#), as well as the particular experiences of C2PA members.

For more information on users, see the Expected Users in the [Guiding Principles](#). Note that this list is not intended to limit consideration of other interested parties.

## 3.2. Phase II: Harm Taxonomy and Assessment

It is worth noting that the potential harms identified reflect system-level considerations that may not be relevant for all products using these specifications.

In an effort to establish a common basis for an analysis and to guide internal and now public discussions, the Taskforce proposes some scenarios based on three temporal stages of the development and adoption cycle of the C2PA standard.

### Scenario 1: Initial Adoption

For this scenario, it is assumed that the tool will be deployed by a few key actors across multiple industries. These



actors will be primarily, though not exclusively, members of the C2PA. Some of these early adopters are actors with significant influence over their respective industries, and it is assumed that their example and reach could lead to a scenario of wide adoption.

### **Scenario 2: Wide Adoption**

It is assumed for this scenario that the C2PA standard could be widely used at a global scale, and that it will be a credible reference of the authenticity and provenance of digital assets. In this scenario, it would be more widely used in social media platforms, by a diversity of media producers and be discussed in legislation or regulation. Despite its widespread use, there would continue to be many actors across different industries, vulnerable groups and geographic locations that do not or cannot use the specifications.

### **Scenario 3: Ongoing maintenance**

This scenario crosscuts through the previous two, and reflects the issue of continuous improvement and adaptation of the specification as a response to a dynamic context and threat landscape.

#### **3.2.1. Identified potential harms**

The table presented [here](#) lists the potential harms identified and classifies them under their respective category and type of harm. The results of the assessment reflect considerations from Scenario 1: Initial adoption.

##### *Identified Harms*

A PDF containing the detailed harms that have been identified and their severity levels can be found [here](#).

Note that the identified mitigations are listed in a separate document that is linked to under the 'Due Diligence Actions' section below.

## **3.3. Phase III: Due Diligence Actions**

Three levels of due diligence actions have been identified:

### **Specifications development**

The specifications should reflect considerations from the harms, misuse and abuse assessment, as outlined in the Guiding Principles.

### **Accompanying documentation**

The accompanying documentation should reflect considerations from the harms, misuse and abuse assessment. The documentation includes:

- Guidance for implementers;
- Guidance on user experience;
- Security considerations;
- Guidance for Artificial Intelligence and Machine Learning;

- Explainer aimed for the general public.

### **Non-technical and multilateral harms response actions**

The harms, misuse and abuse assessment has also highlighted the need for continuously monitoring the impact of the specifications, for developing mechanisms to reflect an evolving landscape and addressing unidentified and unmitigated threats and harms.

Other areas for due diligence actions include:

- Multilateral cooperation for the promotion of a diverse C2PA ecosystem, including efforts to resource public-interest implementations;
- Sensitization of the specifications to a broad base of users and stakeholders;
- Promote parallel efforts to ensure compliance around C2PA specs-enabled products.

## Chapter 4. Public Review and Feedback

Recognizing the limitations and biases of C2PA members and to ensure feedback on harms, misuse and abuse scenarios and responses, the Threats and Harms Taskforce has engaged in a focused effort to solicit input from people and groups across the globe that may consider themselves likely to be impacted by the implementation of these specifications. This feedback has centred on communities with lived, practical, professional and technical experience of the impact of similar technologies, as well as communities that are often excluded from technology design and implementation decision-making while also being the most likely to experience potential harms. Some of the areas covered in these sessions included understanding the potential impact of C2PA specifications on efforts to defend and protect human rights, uphold digital and economic rights, combat mis/disinformation, support civic/community/independent media, and more generally, its impact on social and democratic structures.

# Chapter 5. Due Diligence Actions

The table presented [here](#) lists existing and potential mitigations to each of the potential harms identified to date. These mitigations reflect the specifications and its accompanying documents as they are at the moment of their version 1.0 publication. They also include recommendations for non-technical and multilateral harms response actions that will be developed further to reflect findings from the ongoing harms, misuse and abuse assessment.

Note that for a summary of relevant security features, considerations, and for a threats assessment and countermeasures, see [Security Considerations](#)

## *Actions*

A PDF containing the due diligence actions can be found [here](#).

# Chapter 6. Harms considerations for C2PA stakeholders

As part of due diligence, we offer the following considerations for key stakeholders of the C2PA specifications.

## NOTE

The harms considerations below do not include risks from security threats (e.g. attackers aiming to compromise the security of the system). For more information on the threat modeling process see the [Security Considerations](#).

## 6.1. General considerations for content creators

Content Credentials may be used by content creators to provide cryptographically verifiable provenance information for various reasons, including, but not limited to, safeguarding authorship or providing additional signals of trust. The harms, misuse, and abuse assessment laid out in this document have identified a list of potential harms, some of which may impact content creators. In order to avert and mitigate these potential harms, this assessment has informed and will continue to inform the development of the specifications. It is important to note, however, given the vast diversity of usages, stakeholders, and industries that the C2PA specifications could enable, that all identified potential harms may not be addressed at this level, but will need to be taken up by implementers in consideration of their specific circumstances, users and stakeholders.

As of the current version of the specifications, the C2PA issues the following harms considerations for content creators.

### Technical Accessibility Considerations

#### *Implementing the C2PA specifications (creating C2PA-enabled tools)*

The specifications are open, global, and opt-in, and they use open standards for which there are existing libraries in various programming languages across a range of devices and operating systems/environments in order to facilitate the development of tools that can meet the needs of a diverse group of users. In other words, the Technical Working Group of the C2PA understands that, as of version 1.1 of the specifications, there are no technical barriers to the creation of C2PA-enabled tools and services that meet the needs of a diverse group of users, including those that may be using older devices and operating systems, or others that may have a poor internet connection.

The C2PA has also published a guide for implementers that will be continuously updated in order to facilitate a C2PA ecosystem that reflects a broad spectrum of needs and circumstances.

#### *Using C2PA enabled tools and software*

The decisions of implementors of the specifications may still hinder the use of Content Credentials where they are needed the most. For example, there is the possibility that C2PA claim generators will not operate in pirated software. Additionally, some specification-compliant implementations may prefer to add restrictions to their use. For example, a tool may restrict its use to only newer devices or operating systems, despite the recommendations listed in the

To address accessibility concerns related to the use of C2PA enabled tools and software, it is necessary to promote a diverse C2PA ecosystem that caters to all user groups throughout the world. To this end, the C2PA seeks to cooperate with implementers to work towards effective global accessibility for content creators and consumers.

### Privacy and security considerations

The following potential harms may still occur in specific implementations of the C2PA specifications:

- Inadvertent disclosure of information: C2PA claim generators may automatically add, or require to add, information to manifests that may be sensitive. Specification-compliant implementations may intentionally or unintentionally hide this feature. [User experience guidance](#) has been published alongside the specs to offer a clear acknowledgement of creator consent before a C2PA implementation can begin accumulating data.
- Redacted (deleted) information from Content Credentials may still be accessible: If a soft binding lookup is enabled or required by manifest stores, then previous versions of a manifest with sensitive information may be located.
- The use of C2PA-enabled tools and services in adverse legal or political situations may result in human rights violations: The C2PA specifications include features to protect the privacy of users, but this does not preclude the possibility of malicious actors, including potentially state actors, misusing or abusing the system.

## 6.2. General considerations for content consumers (consumers of Content Credentials)

The C2PA specifications do not provide value judgments about the truth or falsehood of digital assets. In other words, the presence of valid manifests does not mean that anything is ‘true’; validated manifests only establish whether the provenance information can be verified as associated with the underlying asset, correctly formed, and free from tampering.

As of version 1.4 of the specifications, the C2PA issues the following comments and considerations for content consumers:

- Digital assets, such as images and videos, can have valid C2PA manifests and still be deemed to be mis or disinformation.
- Content consumers should note that the verification of an active manifest includes validating the signature. An invalid manifest cannot be associated with the signer.
- The fact that any digital asset does not have Content Credentials does not mean that its contents are not to be trusted. This may be especially relevant to note in a scenario where these specifications are widely adopted. In other words, it is important to know that there may be content creators who will have legitimate reasons to not use C2PA enabled tools and services, and their content should not be dismissed or undermined for not being connected to Content Credentials.
- Invalid Content Credentials could mean that either the digital content or the manifest has been tampered with,

though this may not always be the case.

- The [User Experience Guidance](#) is being designed to define best practices for presenting C2PA provenance to consumers. The recommendations strive to describe standard, readily recognizable experiences that provide asset consumers information and history about the content they are consuming, thereby empowering them to understand where it came from and decide how much to trust it.
- If a content consumer trusts a particular signer, then they may be inclined to believe that the information in the Content Credentials is authentic. In other words, a preexisting relationship of trust between a signer and a content consumer is the basis of the C2PA trust model. C2PA specifications are designed to provide signals for content consumers to know that the signer is in fact who they say they are and that the manifest is in fact connected to a particular asset.

## 6.3. General considerations for civic, community, and independent media

The harms, misuse and abuse assessment laid out in this document identifies a list of potential harms, some of which may impact civic, community and independent media. In order to avert and mitigate these potential harms, this assessment has informed and will continue to inform the development of the specifications. It is important to note, however, given the vast diversity of usages, stakeholders, and industries that the C2PA specifications could enable, that all identified potential harms may not be addressed at this level, but will need to be taken up by implementers in consideration of their specific circumstances, users and stakeholders.

### Technical Accessibility considerations

As of version 1.3 of the specifications, the C2PA issues the following harms considerations for civic, community, and independent media that may be interested in 1. creating their own C2PA-enabled tool, 2. in becoming a signer, or 3. in simply using C2PA-enabled tools.

#### *Implementing the C2PA specifications (creating C2PA-enabled tools)*

In order to guarantee accessibility, privacy, and security criteria, civic, community, and independent media may be interested in creating their own C2PA-enabled tools. In this case, although there are technical and financial requirements to be considered, it is worth noting that the specifications have been designed to facilitate implementations to the extent possible: The specifications are open, global, and opt-in, and they use open standards for which there are existing libraries in various programming languages across a range of devices and operating systems/environments in order to facilitate the development of tools that can meet the needs of a diverse group of users.

That said, the C2PA seeks to cooperate with implementers, partners and other stakeholders in order to promote a diverse ecosystem that caters to all user groups throughout the world, including those that actively strive to guarantee accessibility, privacy and security criteria.

The C2PA has also published a [guide for implementers](#) that will be continuously updated in order to facilitate a C2PA ecosystem that reflects a broad spectrum of needs and circumstances.

## *Becoming a signer*

Civic, community, and independent media may be interested in becoming signers in order to have their brand vouching for Content Credentials tied to the digital assets they create and share. By becoming signers, civic, community, and independent media would also be able to determine what information (assertions) are included in the Content Credentials generated.

There are two ways of becoming a signer: either by using credentials issued by a CA or by self-signing a C2PA manifest. C2PA manifests currently makes use of X.509 certificates which allow for independent verification of your identity, thereby adding a layer of trust to the signed manifests. However, the X.509 certificates come at a cost that may not be accessible to all. For those that self-sign a provenance claim, they should note that these may be deemed to be less credible since the certificate is not independently verified.

In some countries, governments may issue digital certificates to all of its citizens. These certificates could be potentially used to sign C2PA manifests. If government control and surveillance is not regulated, or if there are laws meant to attach journalistic identity to media posted online, these certificates may be used to enforce suppression of speech or to persecute journalists if required by claim generators that do not guarantee privacy and confidentiality.

Similarly, certain C2PA claim generators may allow content creators, including civic, community and independent media, to sign manifests with their personal certificates associated with their IDs. Although guidance to allow for anonymity and pseudonymity has been issued, specification-compliant tools may sell information to third-parties, or not follow user experience guidance meant to empower users to retain control of their information.

## *Using C2PA enabled tools and software*

Content Credentials may be used by the civic, community, and independent media to provide cryptographically verifiable provenance information for various reasons, including, but not limited to, safeguarding authorship or providing additional signals of trust.

Technical restraints at the specifications level may still hinder the use of C2PA-enabled technologies where they are needed the most. For example, it is expected that C2PA claim generators will not operate in pirated software. Additionally, some specification-compliant implementations may prefer to add restrictions to their use. For example, a tool may restrict its use to only newer devices or operating systems, despite the recommendations listed in the Guidance for Implementers.

To address accessibility concerns related to the use of C2PA enabled tools and software, it is necessary to promote a diverse C2PA ecosystem that caters to all user groups throughout the world. To this end, the C2PA seeks to cooperate with implementers to work towards effective global accessibility for content creators.

## **Privacy and security considerations for civic, community and independent media**

The following potential harms may still occur in specific implementations of the C2PA specifications:

- Inadvertent disclosure of information: C2PA claim generators may automatically add, or require to add, information to manifests that may be sensitive. Specification-compliant implementations may intentionally or unintentionally hide this feature.



- Redacted (deleted) information from Content Credentials may still be accessible: If soft binding lookup is enabled or required by manifest stores, then previous versions of a manifest with sensitive information may be located.
- The use of C2PA-enabled tools and services in adverse legal or political situations may result in human rights violations: The C2PA specifications include features to protect the privacy of users, but this does not preclude the possibility of malicious actors, including potentially state actors, to misuse or abuse the system.
- Other considerations for civic, community, and independent media\*

Media and news outlets may face additional pressure to authenticate media, both in terms of ensuring media they are sourcing from stringers, ingesting into their archives, and producing and publishing in-house is resilient to falsifications, and in terms of assessing user-generated content, they include in their reporting. The additional resources required to authenticate media may preclude participation.

## 6.4. General considerations for human rights defenders

Video captured by eye witnesses and on-the-ground human rights activists can be instrumental in drawing attention to human rights violations, supporting calls for policy change, and pushing for accountability.

The C2PA, its implementations, and the broader provenance and authenticity ecosystem could have both positive and negative effects on the way that visual or audiovisual evidence of human rights violations are captured and used.

On the one hand, digital assets embedded with Content Credentials could help add a layer of trust so that evidence of human rights violations are not as easily dismissed or undermined. This could be more relevant now and in the future as technologies to create synthetic media are improved, further blurring the divide between what is real and fake. To name an example, Content Credentials could be used to prevent the [Liar's Dividend](#), or the dismissal of real footage by suggesting that it is a deepfake or in other ways manipulated in order to avoid accountability.

On the other hand, the mere existence of a provenance and authenticity ecosystem, bolstered by the C2PA specifications, could result in higher expectations of forensic proof of visual or audiovisual evidence of human rights violations. If this is the case, questions about accessibility and privacy arise, as well as about the requirements to determine authenticity in legal and social scenarios. To put it differently: Who will take the stand? If jurors and judges come to expect higher levels of admissibility of multimedia content, then witnesses could be asked to verify, corroborate, or authenticate multimedia evidence more frequently. Who will determine what is authentic in this case?

There are also questions around the legal use of Content Credentials. Although the C2PA has not been designed to be used in legal procedures, it may still become an element of consideration in certain scenarios, both to add a layer of trust or to dismiss otherwise authentic content.