



Coalition for Content Provenance and Authenticity

C2PA Harms Modelling

1.0, 2021-12-22: Release

Table of Contents

- 1. Harms, Misuse, and Abuse Framework2
- 2. Methodology4
 - 2.1. Phase I: Purposes, Use-cases, Users and Stakeholders4
 - 2.2. Phase II: Harm Taxonomy and Assessment4
 - 2.3. Phase III: Due Diligence Actions5
- 3. Harms, Misuse, and Abuse Initial Assessment6
 - 3.1. Phase I: Purposes, Use-cases, Users and Stakeholders6
 - 3.2. Phase II: Harm Taxonomy and Assessment6
 - 3.3. Phase III: Due Diligence Actions7
- 4. Public Review and Feedback9
- 5. Due Diligence Actions10

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;
- ¥ 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](#).