Coalition for Content Provenance and Authenticity

C2PA Harms Modelling

0.1, 2021-11-18: Draft (0.1)

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 - Overview . . . 6
4. Public Review and Feedback . . . 10
5. Due Diligence Actions . . . 11

Chapter 1. Harms, Misuse, and Abuse Framework

Harms modelling focuses on analysing how a socio-technical system might negatively impact users and other stakeholders. 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 comprehensively 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 comprehensive 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 comprehensive 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, we 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. We 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 Manifest Task Forces within C2PA.

Below we present some of the modifications we made from existing frameworks for technology impact assessment:

Human rights-focused harm taxonomy

We wanted to ensure 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 we modified the harm taxonomy particular to the Microsoft Harms Modelling Framework to reflect these issue intersections, stakeholders, and users. The reader will note that in this iteration, 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 us to consider a range of potential harms/misuses/abuses.

Temporality

We found it 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

In this initial stage of analysis, we have conducted an internal process to understand severity, scale, likelihood, frequency and impact of potential harms. We have done so in consultation with issue experts within the C2PA and based on C2PA member WITNESS's work 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). However, we see this exercise as an important starting point for feedback rather than a final analysis. Further stages of harm analysis will be conducted in consultation with outside groups, particularly from communities with lived, practice and expert knowledge, and those who will be disproportionately impacted by harms and are often most excluded from design. This analysis should be ongoing, considering that the degree of severity, likelihood, and impact will likely change and become more evident after the technology is deployed.

Considering accountability

Acknowledging that ethical analyses and threat modelling processes are sometimes done behind closed doors, we find it important to emphasize that our harm analysis should be inclusive and it should guide future processes related to governance and compliance.

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.

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, we 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 we were attempting to answer 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, we reviewed and adapted Microsoft's taxonomy of harm to better reflect the context and implications of the C2PA standard and inform it via the other frameworks noted above. 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 is 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 will guide this phase are:

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

Chapter 3. Harms, Misuse, and Abuse Initial Assessment - Overview

The C2PA standard is still in the design stage, and reflects a system specification not a specific product, so the potential harms identified thus far reflect system-level considerations that may not be relevant for all products built using these specifications. A more detailed harms, misuse and abuse assessment will be continuously updated and be available for review through this link.

In an effort to establish a common basis for an analysis and to guide our internal and now public discussions, we propose some scenarios based on three temporal stages of the development and adoption cycle of the C2PA standard. The assessment of the identified harms responds to each one of these scenarios.

Scenario 1: Initial Adoption

In this scenario, we assume 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 we assume that their example and authority could lead to a scenario of wide adoption.

Scenario 2: Wide Adoption

We assume 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/cannot use the standard.

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.

The table below lists the identified harms and classifies them under their respective category and type of harm. In the upcoming versions of the specifications, each harm will be analysed to determine its severity, scale, likelihood and frequency, in addition to the disproportionate risks to vulnerable groups globally.

Table 1. Identified Harms (version 0.8)

Category	Type of Harm	Harm
Risk of injury	Emotional or psychological distress; Physical harm	Misinformation and Malinformation

Category	Type of Harm	Harm
Denial of consequential services	Opportunity loss	Educational discrimination
		Digital divide/technological discrimination (1)(2)
		Journalistic Freedom and Independence
		Loss of choice/network and filter bubble
	Economic loss	Devaluation of individual expertise
		Differential pricing for goods and services
		Increased abuse of systems of creative ownerships
		Creative ownership impersonation

Category	Type of Harm	Harm
Infringement on human rights	Dignity loss	Public shaming, malinformation and targeted exposure and harassment
	Liberty loss, discrimination and due process	Augmented Policing and Surveillance
		Loss of Effective Remedy
	Privacy loss	Interference with private life
		Reduction in options for anonymity and pseudonymity
		Never forgotten
	Constraints on Freedom of Expression	Inability to freely and fully develop personality and creative practice
		Enforcement of extralegal or restrictive laws on freedom of expression
	Freedom of Association, Assembly and Movement	Forced association (Requiring participation in the use of technology or surveillance to take part in society)
		Loss of freedom of movement or assembly to navigate the physical or virtual world with desired anonymity
	Environmental impact	High energy consumption

Category	Type of Harm	Harm
Erosion of Social and Democratic Structures	Manipulation	Misinformation
	Over-reliance on systems	Overconfidence in technical signals
	Amplification of power inequality	Social detriment
		Journalistic plurality and diversity

Chapter 4. Public Review and Feedback

Recognizing the limitations and biases of C2PA members and to ensure feedback on harm, misuse and abuse scenarios and responses, we now solicit input from as many voices as possible, particularly from people and groups across the globe that may consider themselves likely to be impacted by the implementation of this standard. This feedback should center on communities with lived, practical and technical experience of the impact of similar technologies, as well as communities most likely to experience potential harms and that are often excluded from technology design and implementation decision-making.

Chapter 5. Due Diligence Actions

The aim of due diligence actions is to mitigate potential abuse and misuse, to offer considerations and guidelines for the protection of human rights, to identify compliance or non-compliance standards, and ensure the optimization of the benefits in terms of trust in media, user control and transparency that prompted the development of the C2PA standard.

Due diligence actions will be developed parallel to the public review and feedback process.