C2PA Harm, Misuse, and Abuse Assessment

PHASE II - Initial Adoption *(Version 0.8 - October 28, 2021)*

Scenarios / Assumptions for assessment

The Harms, Misuse, and Abuse Assessment cannot be exhaustive. In an effort to establish a common basis for an analysis and to guide internal and public conversations, we propose some scenarios / assumptions based on three stages of the tool adoption/development.

1. Initial Adoption: The standard is adopted by a few key actors across various industries (C2PA members, primarily)

2. Wide Adoption: The C2PA standard is widely used at a global scale as a credible reference of the authenticity and provenance of digital assets.

3. Ongoing maintenance / response: The standard is continuously improved to respond to the context in which C2PA technology is deployed and to changes in the threat landscape.

The harms, misuse, and abuse assessment is an ongoing process. The information presented below should not be considered the end result of a comprehensive evaluation, but a basis for broader, ongoing, and more profound discussions, centering on impacted communities, that could lead towards the mitigation of potential harms, abuse and misuse and the protection of human rights.

To address the harms listed below, and to identify other potential harms, C2PA members are leading conversations with a broad groups of stakeholders.

The aim of these efforts is to shape 1. the specifications and their development process, 2. guidance documents for implementers, 3. governance models, so that identified and emerging harms may be addressed.

We encourage stakeholders to participate in identifying potential harms, and developing potential mitigation strategies.

Category	Type of Harm	Potential Harm / Misuse / Abuse	Contextual Example / Evidence	Existing and Potential Mitigations
		Language discrimination Limited language versioning on C2PA-enabled tools, despite their focus on low-cost and global accessibility, leads to more limited access for marginal markets.	C2PA-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.	**Ongoing identification of existing and potential mitigations.**
		Digital divide/technological discrimination (1) 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 a customer can access.	For example, existing experiences with gated/limited access to particular websites and tools via Free Basics program for "affordable access" from Facebook / mobile operators in emerging markets. See also example above on **Educational discrimination** and limited language versioning.	

Denial of consequential services	Opportunity loss	**Digital divide/technological discrimination (2)** Individuals and communities without abiilty to access or use tools for compliance with system usage are excluded.	Financial costs involved in signing up to use different C2PA-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.	potential mitigations.
		Journalistic Freedom and Independence 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. Misuse of provenance datastores to track content or enforce restrictive laws on freedom of expression and do so with lack of effective remedy and/or exploitation of provenance datastores to track content, and curtail freedom of expression (e.g. political speech). See overlap with **Journalistic Plurality and Diversity**	An escalation of laws addressing 'fake news', misinformation/disinformation and social media globally (c.f. statistic from ICNL) includes laws that enforce registered identity as a journalist on social media (e.g. Tanzania), provide governmental right-to-reply (Singapore), are being used to suppress dissent and reduce journalistic freedom.	**Ongoing identification of existing and potential mitigations.**
		Loss of choice/network and filter bubble Risk of exacerbating epistemic injustice (whose accounts and knowledge are heard, validated and trusted) reflecting power dynamics of access and privilege among consumers and producers (including media) in terms of which information gets C2PA signals.	Existing dynamics of how both accreditation systems and epistemic trust focus on professional experience and exclude non-professional, community, non-accredited and historically marginalized communities.	
	Economic loss	**Devaluation of individual expertise** C2PA-based technology displaces skilled fact-checkers/journalists from news organizations and civil society watchdogs.	Existing precedents of displacement/shifts in employment of a range of skilled and unskilled professions by automation.	
		Differential pricing for goods and services Price discrimination as a result of participation in the marketplace for creative content or journalistic content disproportionately excludes marginalized communities and non-mainstream media who do not have access to relevant tools, or cannot consistently use tools because of privacy or other reasons. See overlap with **Journalistic Plurality and Diversity; Digital Divide, Forced association (Requiring participation in the use of technology or surveillance to take part in society)**	Existing precedents of increased pricing for content with particular characteristics perceived as valuable in marketplace	
		Increased abuse of systems of creative ownerships C2PA-enabled attribution supports more extensive copyright trolling based on analysis of C2PA data.	Existing precedents of copyright trolling. Eg. U.S. Righthaven LLC suing DiBiase, an attorney who leads nobodycases.com in order to provides resources for difficult-to-prosecute murder cases, for copyright infringement. See Righthaven v. DiBiase. See also Higbee and Associates suing Homeless United for Friendship and Freedom ("HUFF") for republishing a NYTimes articles that was featured on their website along with a copyrighted image.	

		Creative ownership impersonation C2PA assertions are used to make an apparent claim of ownership on another creator's work.	C2PA-backed assets may be used to claim ownership, c. f. a corporate NFT actor decided to create NFTs from CC licensed museum images at the Rijksmuseum in the Netherlands.	
	Dignity loss	**Public shaming, malinformation and targeted exposure and harassment** This may mean exposing people's private, sensitive, or socially inappropriate material (for example via doxxing based on C2PA-derived data, or using media created with C2PA data). See overlap with **Interference with Private Life** and **Never Forgotten**	Taking current or historical sensitive data from online platforms/services/devices and using C2PA to target particular groups such as women or women's right groups/LGBTQIA+ groups based on traffic. This can be both indiviual data and aggregate data. For example: activists where LGBTQIA+ is criminalized being deanonymized from individual and aggregate Grindr data to reveal use of the app.	
	Liberty loss, discrimination and due process	**Augmented Policing and Surveillance (1)** If data from C2PA and other sources were to be aggregated, it could be used to target and discriminate against individuals and groups. This could occur for example with police body cams if they are equipped with facial recognition technologies that reinforce racial and other biases. C2PA data is used to amplify surveillance mechanisms and to infer suspicious behavior and/or criminal intent based on historical records.	Harms around facial recognition are now well documented in a law enforcement context. Data from C2PA could also be incorporated into invasive online identification schemes.	
		Augmented Policing and Surveillance (2) Biometric identification approaches incorporate C2PA-spec'd devices for capture and enhanced protection for biometric scans for identification, resulting in additional data for identification of individuals, with potential privacy-compromising and impacts on both obligatory usage for marginalized-community as well as exclusion consequences.	Biometric and digital identity systems deployed without public accountability	
		Augmented Policing and Surveillance (3) Privacy loss for consumers of media via tracking of access to cloud-based assertion data (e.g. via tracking pixels)	Eg. tracking pixels of access to cloud-based assertion data.	
		Augmented Policing and Surveillance (4) Use of broader availability of provenance data to do broad search for content, e.g. geofenced location data search	Significant concern since journalistic/media usage of a C2PA-enabled search is an identified use case	
		Loss of Effective Remedy (1) Usage of C2PA signals in automated systems lack capacity to adequately interpret a complex set of complementary signals. Consumers lack transparency and right of appeal to potential algorithmic bias or harms from interpretation of C2PA signals.	Loss of effective remedy in existing systems of algorithmic recommendation or downranking, or algorithmic content removal. In C2PA, decoupled, hard bound manifests are automatically matched to its content. Although unlikely, hash collitions cannot be ruled out. C.f. literature on accuracy of automated systems in complex contextual situations - e.g. misinformation ("Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis", Center for Democracy and Technology") and on opacity in authenticity infrastructure	
		Loss of Effective Remedy (2) In case of an inaccurate or misleading C2PA result, individuals will not have the ability to contest a technical decision that may have repercussions on a personal or community level	Existing forensic explainability challenges of media forensics - c.f. the controversy even over whether/how a World Press Photo prize-winning photo was manipulated.	

Infringement on human rights	Privacy loss	**Interference with private life (1)** Inadvertent disclosure of information (from unintended inclusion of assertions, or disclosure of assertions in an unintended way) or aggregate information from assertion combined with other data. See overlap with **Never Forgotten** and **Public Shaming**	Taking sensitive data from online platforms/services/devices and using C2PA to target particular groups such as women or women's right groups/LGBTQIA+ groups based on traffic. For example: activists where LGBTQIA+ is criminalized being deanonymized from Grindr data to reveal use of the app.	**Ongoing identification of existing and potential mitigations.**
		Interference with private life (2) Misuse of C2PA soft-binding for broadened scope of data-rich searchable hash stores outside of traditional 'violating' hash databases including use in emergent client-side data scanning in encrypted messaging.	Recent experience with Apple's plans to scan client-side photos for Child Sexual Abuse Material (CSAM). Although it was meant to ensure child safety, it had broad and potentially harmful implications. See International Coalition Calls on Apple to Abandon Plan to Build Surveillance Capabilities. Existing client-side scanning in PRC	
		Reduction in options for anonymity and pseudonymity Inadvertent disclosure of information (from unintended inclusion of assertions, or disclosure of assertions in an unintended way) or aggregate information from assertion combined with other data.	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) Aggregate mobile data provides the ability to deanonymize individuals, for example through phone reversal lookup APIs, or through public records search or breach in the case of countries where biometric data is required for telecommunication services (c.f. Mexico reform to the federal telecommunications law).	
		Never forgotten Digital files or records may never be deleted. Depending on a given C2PA-enabled system's functionality for redaction of soft or hard binding provenance datastores, and/or what usage of irrevocable ledgers, risk of digital files that contain misinformation OR that contain privacy and dignity-compromising information continues to circulate. Storage of manifests may not allow for manifest redaction, deletion or selective disclosure. A content creator may want to delete private/sensitive manifest from content, but decoupled manifests in provencnace data stores may reattach manifest to content (c.f. analogous to Google/Facebook image storage). Further questions around the capacity to redact information in blockchain-based C2PA systems remain.	For example: human rights defenders, journalists and others in Afghanistan removing content showing their face and personal information from Internet/social media	
	Constraints on Freedom of Expression	**Inability to freely and fully develop personality and creative practice** Workplace requirements to use tools for production in journalistic/creative contexts may have implications for personal privacy and personal artistic practice by forcing disclosure of techniques	Media and artistic/creative producers are concerned about disclosure of creative techniques	
		Enforcement of extralegal or restrictive laws on freedom of expression Misuse of provenance datastores to track content or enforce restrictive laws on freedom of expression and do so with lack of effective remedy and/or exploitation of provenance datastores to track content, and curtail freedom of expression (e.g. political speech)	Political dissidents being tracked through C2PA datastores, or 'bad actors' demanding datastores to release sensitive information.	
	Freedom of Association, Assembly and	**Forced association (Requiring participation in the use of technology or surveillance to take part in society)** De facto inclusion and participation obligation in marketplaces for creative content or journalistic content or for better algorithmic ranking on social media sites which disproportionately excludes global populations, marginalized communities and non-mainstream media who do not have access to relevant tools, or cannot consistently use tools because of privacy or other reasons.	For example, algorithmic ranking: content creators forced to game algorithms with particular keywords, metadata to achieve visibility/to be ranked higher in a feed.	

	Movement	**Loss of freedom of movement or assembly to navigate the physical or virtual world with desired anonymity** C2PA-enabled systems that utilize a real-name identity or other real-world profile provide a mechanism to connect movement in space to an individual via C2PA metadata	Inadequate UX or implementation creates simplistic signals of trust that obscure real-life dynamics faced by individuals who mix some elements/moments of public visibility with pseudonymity and anonymity in other circumstances.	
	Environmental impact	**High energy consumption** Extensive use of blockchain or types of distributed ledger technology with C2PA-enabled systems contributes to exploitation of natural resources.	Blockchain-enabled C2PA systems would be part of a broader high-energy consumption ecosystem. The assessment however reflects the assumption that these systems would operate with proof-of-stake models.	
Erosion of Social and Democratic Structures	**Manipulation**	**Misinformation (1)** C2PA-enabled systems can be used to generate misinformation (for example by generation of deliberately misleading manifests) and imply that it is trusted.	*[From security considerations]* An attacker misuses a legitimate claim generator (e.g. C2PA-enabled photo editor) to add misleading provenance to a C2PA-enabled media asset. See Journalistic plurality and diversity	**Ongoing identification of existing and potential mitigations.**
		Misinformation (2) C2PA-enabled tools can be used to support misinformation, and in certain circumstances make it hard to revoke or retract this misinformation, leaving a continued assumption of additional trust.	An erroneous C2PA manifest is used to "validate" an image that is widely shared. C2PA manifest is not revoked or retracted, so the image is continuouslly shared and trusted.	
		Misinformation (3) Disguising fake information as legitimate or credible information by deliberate mis-attribution and assignation of C2PA provenance to existing content (without C2PA data) and legacy media, and addition of relevant soft and hard bindings to provenance datastores where look-up provides deceptive results on first visual glance.	For example, if using thumbnails or other low quality images to do a soft binding look up of an asset throws back a wrong match. This information could then be used to misinform. C2PA validator fosters a loss of remedy in cases like these.	
		Misinformation (4) C2PA-enabled ecosystem creates an 'implied falsehood' around media that does not contain C2PA assertions/manifests, resulting in discrediting of legitimate content sources.	Videos or images from sources that cannot or prefer to not used C2PA-enabled devices are discredited or undermined.	
		Misinformation (5) Mislabeling trustworthy information as misinformation: C2PA-enabled tools and derived signals AND/OR soft-binding hashes can be used inappropriately in automated systems for detecting, classifying, organizing, managing and presenting misinformation.	C.f. analysis on automation of visual misinformation detection	
	Over-reliance on systems	**Overconfidence in technical signals** Over-confidence in the technical signals as an indicator of truth or confirmation of trust, rather than a set of signals related to provenance and authenticity/edits. Use of automated look-up systems progressively reduces human-in-the-loop, leading to exacerbated problems around contextualization of information or augmentation of problems (for example, mislabeling of misinformation based on contextual misunderstanding, or from malicious uses articulated above). These problems could occur at the front-end providing deceptive UX assumptions to soft-binding look-up or at back-end with over-automation of usage of soft-binding signals. See overlap with **Loss of remedy** and automation	C.f. literature on overconfidence in simple technical signals, particularly in misinformation systems. In C2PA specifications, an open question remains on the issue of automatically linking digital assets to manifests in provenance databases through soft binding matches.	

	Social detriment	**Amplification of power inequality** Requiring participation in the use of technology to take part in society.	De facto inclusion and participation obligation in marketplaces for creative content or journalistic content or for better algorithmic ranking on social media sites which disproportionately excludes global populations, marginalized communities and non-mainstream media who do not have access to relevant tools, or cannot consistently use tools because of privacy or other reasons.	
		Journalistic plurality and diversity "A divergence of usage between media able to afford/adapt to/use C2PA-enabled tools and workflows, and a broader range of smaller media and individual citizen journalists leads to a de facto two tier trust system in public perception." See overlap with **Journalistic Freedom and Independence**	Smaller or community news publishers are unable to provide C2PA-backed content, and so their content is undermined by audiences, platforms, governments, influential individuals (eg. Liar's dividend).	
Risk of injury	**Physical or infrastructure damage**	*No identified harm; pending consultation with other stakeholders.*	*No identified harm; pending consultation with other stakeholders.*	**Ongoing identification of existing and potential mitigations.**
	Emotional or psychological distress; Physical harm	**Misattribution and Malinformation (2)** Misuse of C2PA-enabled media to implicate an individual or group in inciting violence/criminality, or otherwise negatively or positively impact the reputation of a group or individual or media entity. Including deliberate mis-attribution and assignation of C2PA provenance to existing content (without C2PA data) and legacy media, and addition of relevant soft and hard bindings to provenance datastores where look-up provides deceptive results on first visual glance (e.g. thumbnail approach) See overlap with **Never Forgotten**	Existing problems of digital wildfire (rapidly-shared online content) frequently feature existing shallowfaked, miscontextualized content claimed to be from one place when actually from another. Patterns of manipulating media to misattribute are commonplace and should be assumed as an attack vector for C2PA-enabled systems.	