HAW HAMBURG

Fakultät Design, Medien und Information Department Medientechnik

Bachelorarbeit

Annika Herzog
Matr.-Nr. [redacted]

Authentifizierung digitaler Bilder durch Metadaten und Bildforensik vor dem Hintergrund KI-generierter Bilder und Citizen Media im Kontext redaktioneller Arbeit

Können C2PA Metadaten (Content Credentials) auf lange Sicht die Authentifizierung von digitalen Bildern für Redaktionen und Rezipient:innen erleichtern?

Erstprüfer: Prof. Dr. Marco Grimm
Zweitprüferin: Nathalie Mai

Version 1 vom 2. Juni 2025
Eingereicht am

Zusammenfassung

Diese Bachelorarbeit untersucht Metadaten in digitalen Bildern im Kontext der neuen C2PA-Spezifikation. Ziel ist es, die Funktionsweise, das Potenzial sowie Risiken und Schwächen der Spezifikation zu analysieren und die Belastbarkeit von Metadaten als Indiz für Bildauthentizität - insbesondere in der Bildforensik und im Journalismus sowie in Citizen Media- zu bewerten. Neben einer Literatur- und Online-Recherche wurden praktische Versuche durchgeführt, um C2PA-Implementationen zu testen und Aussagen aus der Recherche nachzuvollziehen. Die Ergebnisse zeigen, dass Metadaten für die Authentifizierung und Verifikation von Bildern entscheidend sein können. Die C2PA-Spezifikation schlägt eine einheitliche Kennzeichnung von KI-generierten Bildern vor, und ermöglicht das Einbinden von Herkunftsnachweisen (Content Credentials), die den Entstehungsprozess eines Bildes nachvollziehbar machen. Die C2PA-Metadaten können einfach gelöscht werden und ihre Implementation ist von externer Infrastruktur abhängig. Für die Akzeptanz und Wirksamkeit der Spezifikation ist neben der möglichst fehlerfreien, technischen Umsetzung insbesondere die verständliche Darstellung der Content Credentials und die Aufklärung der Nutzer:innen entscheidend. Insgesamt zeigt die Arbeit, dass Metadaten eine Schlüsselrolle in Verifikationsprozessen spielen, der Begriff „Authentizität“ vor Umsetzung der C2PA-Spezifikation jedoch differenzierter betrachtet werden muss.

Abstract

This bachelor's thesis examines metadata in digital images within the context of the new C2PA specification. The aim is to analyze the functionality, potential, as well as risks and weaknesses of the specification and to evaluate the reliability of metadata as an indicator of image authenticity—especially in image forensics, journalism, and citizen media. In addition to literature and online research, practical experiments were conducted to test C2PA implementations and verify statements found throug research . The results show that metadata can be crucial for authenticating and verifying images. The C2PA specification proposes a unified way of labeling AI-generated images and enables the integration of provenance information (Content Credentials) that make the image creation process traceable. However, C2PA metadata can be easily deleted, and its implementation depends on external infrastructure. For the acceptance and effectiveness of the specification, besides technically robust implementations, clear presentation of Content Credentials and user education are essential. Overall, the study demonstrates that metadata play a key role in verification processes, but the concept of "authenticity" needs to be viewed with greater complexity before applying the C2PA specification.

Inhaltsverzeichnis

Abkürzungsverzeichnis

Adobe	Adobe Inc.
AIGC	with Artifical Intelligence Generated Content
AP	Associated Press
APP	Application Marker Segments
BBC	British Broadcasting Corporation
BSI	Bundesamt für Sicherheit in der Informationstechnik
C2PA	Coalition for Content Provenance and Authenticity
CAI	Content Authenticity Initiative
CBC	Canadian Broadcasting Corporation
CBOR	Concise Binary Object Representation
CCr	Content Credentials
DLT	Distributed Ledger Technology
dpa	Deutsche Presseagentur
DW	Deutsche Welle
EBU	European Broadcast Union
Exif	Exchangable Image File Format
GIMP	GNU Image Manipulation Program
IFD	Image File Directory

IPTC	International Press Telecommunications Council
ITI	Information Technology Industry Council
ITU	Internationale Fernmeldeunion
ITU-T	ITU Telecommunication Standardization Sector
NAA	Newspaper Association of America
NIST	National Institute of Standards and Technology (USA)
NYT	New York Times
OSINT	Open Source Intelligence
PKI	Public Key Infrastructure
PRNU	Photo Response Non-Uniformity Pattern
PS	Photoshop, von Adobe Inc.
RDF	Resource Description Framework
SWGDE	Scientific Working Group on Digital Evidence
TIFF	Tagged Image File Format
UGC	User Generated Content
WDR	Westdeutscher Rundfunk
XML	Extensible Markup Language
XMP	Extensible Metadata Platform

Abbildungsverzeichnis

1 Einleitung

Die Macht der Bilder kombiniert mit den immer realistischer werdenden KI-generierten Inhalten fordert Zeitungsverlage, Nachrichtenagenturen und Bildredaktionen heraus. 2020 schrieb eine Gruppe von Microsoft-Assoziierten in einer Vorveröffentlichung: „we expect that the march of technical progress will soon make it impossible to distinguish fake media from real“ (England u. a. 2020). Zwei Jahre später wird evaluiert, dass generative KI das *uncanny valley* verlassen hat, und Gesichter[1] generiert, die nicht mehr von realen zu unterscheiden sind (vgl. Nightingale u. a. 2022; SWGDE 2025, S. 8).
Bilder werden im journalistischen Kontext genutzt, um Wahrheit zu vermitteln (vgl. Gerling 2022), haben damit Auswirkungen auf die Meinungsbildung und demokratische Prozesse. Werden Unwahrheiten verbreitet, hat das kurz-, mittel- und langfristig weitreichende Konsequenzen: Die Glaubwürdigkeit der Medien[2] leidet, zusammen mit anderen Technologien[3] wird die Beeinflussung einer Bevölkerung erleichtert, letztlich könnten Demokratien dadurch geschwächt werden[4]. Obwohl die digitale Fotografie und mit ihr die digitale Bildbearbeitung keine neue Technologie ist, glauben viele Menschen weiterhin unhinterfragt den Aussagen eines Bildes. Umso größer ist die Verantwortung der publizierenden Unternehmen, Bilder vor Veröffentlichung zu authentifizieren und zu verifizieren.

[1] Für eine Kostprobe empfiehlt sich die Seite `https://thispersonnotexist.org/`.
[2] In diesem Fall sind mit dem Begriff Medien die Medienanstalten, News Outlets und mittelbar auch Social-Media-Plattformen gemeint.
[3] Etwa Deepfakes, automatisierte, KI-gestützte Gesichtserkennung und KI im Allgemeinen.
[4] Das untersucht zum Beispiel das Bundeskriminalamt in seiner Literaturauswahl zur BKA Herbsttagung im November 2024 (vgl. Bundeskriminalamt 2024).

Fragestellung und Ziel dieser Arbeit

Wie lässt sich ein KI-generiertes, oder auch nur mittels KI in seiner Aussage manipuliertes Bild, von einer zwar bearbeiteten aber dennoch authentischen[5] digitalen Fotografie unterscheiden?
Metadaten können bei der Suche nach einer Antwort helfen. KI-generierte Bilder enthielten zu Beginn der Entwicklung nur wenige Metadaten, das möchte die Content Authenticity Initiative (CAI) zusammen mit der Coalition for Content Provenance and Authenticity (C2PA) ändern. Die C2PA-Spezifikation soll Metadaten kryptografisch an die Bilddaten (*image data*) binden, und so fälschungssicher speichern; sowohl in digitalen Fotografien als auch in synthetischen Bildern. Die Interpretation der C2PA-Metadaten soll in Form von Content Credentials (CCr) sichtbar gemacht werden, sodass sowohl Rezipient:innen als auch Journalist:innen die Herkunft des Bildinhaltes beurteilen können, ohne das Bild auf technischer Ebene analysieren zu müssen.
Diese Arbeit verfolgt ein zweigeteiltes Ziel. Es soll erstens untersucht werden, wie belastbar die aus Metadaten gewonnenen Informationen bei der Beurteilung der Authentizität eines Bildes sind. Zweitens soll erörtert werden, welchen Einfluss die systemweite Implementierung der C2PA-Spezifikation auf diese Belastbarkeit und das Mediensystem insgesamt haben kann.

Können C2PA-Metadaten (Content Credentials) auf lange Sicht die Authentifizierung von digitalen Bildern für Redaktionen und Rezipient:innen erleichtern?

Um Desinformation entgegenzuwirken, müssen Bilder *vor* Veröffentlichung verifiziert werden können. Deswegen ist alles, was diesen Verifikationsprozess vereinfacht, ein Teil der Lösung. Authentifizierung ist ein Teil der Verifikation. Die Verifikation ist eine Vorgehensweise, bei der die Glaubwürdigkeit von Informationen beurteilt werden soll. Im journalistischen Kontext findet sie *vor* der Veröffentlichung der Information statt, bezieht Kontextinformationen mit ein, und konzentriert sich auf die Verlässlichkeit der

[5] Eine dem Kontext angemessene Betrachtung des Begriffs erfolgt in Abschnitt 5.1.

Quelle und Richtigkeit ihrer Aussage. Die Authentifizierung von Bildern hingegen nutzt nur die Bilddatei allein, um zu einer Einschätzung zu gelangen. Die Implementierung von Content Credentials auf Social-Media-Plattformen könnte außerdem dazu führen, dass Nutzer:innen proaktiv werden und Bild- und andere Inhalte eher hinterfragen. Citizen Media[6] könnten durch Content Credentials einfacher verifiziert werden und damit häufiger Einzug in die öffentliche Berichterstattung erhalten. Weiterhin würden durch Content Credentials KI-generierte Bilder[7] als solche identifiziert werden können, auch wenn es keine visuellen Anhaltspunkte für den synthetischen Ursprung gibt.

Einschränkungen

Metadaten gibt es in Text, Bild, Video und Audio. Diese Arbeit beschränkt sich auf digitale Bilder, wenngleich vieles Erarbeitete genauso für digitale Videos gilt. Der Fokus liegt auf journalistischen Bildern, die im Nachrichten-Kontext und auf Social Media zur Verbreitung von Nachrichten verwendet werden.

Durch die vielen, parallel verwendeten Standards gibt es eine große Anzahl an möglichen Metadaten-Einträgen in Bildern; es werden markante[8] Metadaten-Felder in den Versuchen untersucht, bearbeitet und analysiert.

Das Forschungsfeld der Bildforensik ist groß. Diese Arbeit konzentriert sich auf die Rolle von Metadaten innerhalb der Bildforensik. Andere Methoden werden ergänzend genannt und kurz beschrieben. Bildredaktionen und Faktencheck-Teams nutzen sowohl bildforensische Methoden als auch Open Source Intelligence (OSINT) in ihren Verifikationsprozessen. Die ausführliche Beleuchtung von OSINT Methoden überstiege jedoch den Rahmen dieser Ausarbeitung.

Die tatsächliche Bedeutung von Metadaten innerhalb journalistischer Praktiken wird nicht empirisch nachgewiesen, hierfür sei auf andere Bachelor-und Masterarbeiten verwiesen[9]. Es erfolgt jedoch eine Einordnung, welche Bedeutung ihnen im erwähnten

[6] Citizen Media bezeichnet erstellte Inhalte von Menschenrechtsaktivist:innen und Journalist:innen der Menschenrechtsszene, oder Aufnahmen von unbeteiligten, zufälligen Zeugen Okeowo 2022; Gerling 2022, S. 22. Stark verallgemeinert wird online auch von User Generated Content (UGC) gesprochen.

[7] Es wird synonym bei KI-generierten Bildern auch von synthetischen Bildern gesprochen.

[8] Die Auswahl erfolgt begründet.

[9] Beispielsweise Zettelmeister 2024, und Peters 2023

Rahmen zukommen *könnte*.
Die C2PA-Spezifikation ist noch in Arbeit. Die hier betrachtete Version 2.1 wurde im September 2024 veröffentlicht, die Beschreibung in Abschnitt 2.2 und Bewertung in Kapitel 5 ist entsprechend als Momentaufnahme zu lesen.

Vorgehensweise

Kapitel 2 beleuchtet die in Bildern vorhandenen Metadaten und die dahinter liegenden Standards. Besonderes Augenmerk liegt auf der C2PA-Spezifikation. Deren Intention und Funktionsweise werden dargestellt. In Kapitel 3 werden die praktischen Versuche beschrieben und ausgewertet. Bildforensische Methoden werden in Kapitel 4 erläutert. Der Fokus liegt auf Methoden, die auch im Journalismus angewendet werden (können). Kapitel 5 führt die theoretischen und praktischen Erkenntnisse über die C2PA-Spezifikation zusammen; nachdem zunächst der Begriff der Authentizität näher betrachtet wurde, erfolgt eine kritische Auseinandersetzung mit der C2PA-Spezifikation und ein Blick auf aktuelle Implementationen. Die Synthese der theoretischen Betrachtungen, praktischen Erkenntnisse und der kritischen Analyse erfolgt in Kapitel 6, sodass abschließend die Forschungsfrage beantwortet werden kann.

Literaturrecherche

Die Arbeit stützt sich zu einem großen Teil auf Internet-Recherchen, weil die C2PA-Spezifikation erst vor wenigen Jahren (2022) veröffentlicht wurde. Einige wissenschaftliche Quellen erwähnen die Idee der C2PA zwar, doch die Auseinandersetzung mit deren Umsetzung erfolgt überwiegend in Online Artikeln, Blogs und Webseiten von beteiligten, oder betroffenen Unternehmen und Akteuren. Zu den Themen Bildforensik, journalistische Praxis und Citizen Media sowie zum Umgang mit dem Begriff Authentizität gibt es sowohl ältere Buch-Literatur vom Beginn des Jahrtausends, als auch jüngere Beiträge und wissenschaftliche Erscheinungen. Die Autorin ist bemüht, nicht-wissenschaftliche Quellen als solche einzuordnen.
Zunächst erfolgte eine freie Recherche zum übergeordneten Thema „Metadaten in Bil-

dern“, der C2PA und ihrer Spezifikation, im Anschluss wurde gezielt nach Kritik, Anwendungsbeispielen und Einsatzzwecken zur Spezifikation gesucht. Die Themenfelder Bildforensik und journalistische Praxis wurden ergänzend erschlossen, um der Komplexität des Themas Rechnung zu tragen. In Form von Versuchen wurden die während der Recherchen aufgekommenen Fragen und Behauptungen untersucht. Hierbei spielten die Arbeiten von Neal Krawetz eine bedeutende Rolle. Kein Anderer setzt sich öffentlich[10] so intensiv und kritisch mit der C2PA auseinander. Andere kritische Nutzer:innen stützen seine Aussagen, auch wenn Krawetz nicht unumstritten ist.

Verwendete Hilfsmittel

Die für die Versuche genutzte Software und Hardware ist im Anhang aufgelistet und erläutert.
Für eine steilere Lernkurve im Umgang mit Software (Exiftool, c2patool, Adobe Photoshop sowie LaTex, Zotero und Obsidian) wurde je nach Verfügbarkeit mit kostenlosem Zugang der KI-chatBot chatGPT 3.5 oder chatGPT 4.0 verwendet. Zum Teil hat die KI beim Verständnis und der Übersetzung englischer Quellen unterstützt; Übersetzungsvorschläge wurden jedoch nie ohne Prüfung durch Wörterbücher und eigene Kontextanalyse übernommen. Die Übersetzung der Zusammenfassung (*abstract*) wurde von Perplexity vorgenommen.

[10] In Form von Blogbeiträgen auf seiner Seite hackerfactor.com.

2 Alte und neue Metadaten

Menge und Inhalt beziehungsweise Informationsgehalt von Metadaten in Bildern waren von jeher an den Verwendungszweck gekoppelt. Allgemein dienen Metadaten dem „Management von gespeicherten Nutzdaten“ (Klussmann 2000, S. 483ff). Sie sind die „systeminternen Daten, die zur Verwaltung der eigentlichen Nutzdaten verwendet werden“ (ebd.). Die Speicherstruktur ist dabei vom Anwendungsfeld abhängig. Anders als in Bibliothekskatalogen, in denen Metadaten zu Büchern getrennt von diesen aufbewahrt werden, und in Form einer Signatur einen Verweis auf den eigentlichen Standort des Buches enthalten, werden Daten über Bilddaten zusammen mit diesen in eine Datei geschrieben. Metadaten-Standards sind das Resultat der Bemühungen unterschiedlicher Akteure im Verlauf der Zeit. Im Bezug auf digitale Bilder haben sich mit der Aufnahme- und Verarbeitungstechnologie auch die Ansprüche an die in ihnen enthaltenen Metadaten entwickelt.

2.1 Etablierte Standards

Das Tagged Image File Format (TIFF) ist ein Dateiformat für gerasterte Bilder (vgl. Klussmann 2000, S. 755). Die von Adobe 1995 zuletzt veröffentlichte Spezifikation[1] beschreibt, dass jede Information in einem Image File Directory (IFD) Eintrag abgelegt werden muss. Eine TIFF-Datei besteht aus einem Header, einem IFD mit mehreren Einträgen, und den eigentlichen Bilddaten. Später entwickelte Standards bauen auf diese

[1] Ursprünglich wurde diese in der Version 3.0 vom Unternehmen Aldus veröffentlicht (vgl. Klussmann 2000, S. 755). Das Unternehmen war auf Publishing Software spezialisiert und wurde zehn Jahre nach Gründung von Adobe aufgekauft. Die Spezifikation wurde als ISO 12639:2004 festgehalten und ist somit nicht mehr frei zugänglich.

Struktur auf.
Mit dem Exchangable Image File Format (Exif) erweiterte sich die mögliche Anzahl der IFDs (Nummerierung startet mit 0) und schafft so Platz für anwendungsbezogene Metadaten, die in ihnen zugewiesenen Application Marker Segments (APP) gespeichert werden. Für die weiteren Betrachtungen sind unter Anderem die Exif-Einträge `FileSource` und `SceneType` interessant. Die Zeitstempel in Exif sind ebenfalls von Bedeutung, wenngleich falsche Einstellungen in der Kamera auch unzutreffende Zeitstempel in den Metadaten erzeugen. Die International Press Telecommunications Council (IPTC) ist eine Organisation zur Standardisierung und Interessensvertretung von Telekommunikations- und Nachrichtenunternehmen. 1991 veröffentlichte sie zusammen mit der Newspaper Association of America (NAA) den IPTC-IIM Standard, um die Übertragung von (Meta-) Daten zu vereinheitlichen und so die korrekte Verwendung der Bildersowohl in Produktion als auch Vertrieb zu ermöglichen. Mit der voranschreitenden Entwicklung der Extensible Markup Language (XML) und Adobes Bestreben, eine eigene Struktur für das Verwalten von Metadaten einzuführen, wurde der Standard 1997 eingefroren. Die Felder wurden jedoch in die IPTC Photo Metadata und IPTC Video Metadata Standards übernommen und können auch mit XMP dargestellt werden. Mittels IPTC-Einträgen können Informationen wie `CopyrightNotice`, `SpecialInstructions` sowie `Keywords`, `AltTextAccessibility` für alternative Bildbeschreibungen und `SceneCode` für Szenenbeschreibungen hinzugefügt werden (vgl. Abschnitt 3.2) .
Während der Wert eines Feldes in Exif größtenteils von den Kameraherstellern bestimmt wird, existiert für viele IPTC-Einträge ein *Namespace*, der vorgibt, welche Werte genutzt werden dürfen, und was sie bedeuten. Festgehalten werden diese Vorgaben im NewsCode[2]. Im Gegensatz zu Exif-Einträgen werden IPTC-Einträge manuell gefüllt.
Die Extensible Metadata Platform (XMP) wurde 2001 von Adobe erstmals veröffentlicht. Ziel war es, bestehende Metadatenstandards zu vereinheitlichen. Mit XML als Sprache und Resource Description Framework (RDF) als Struktur werden Metadaten maschinenlesbar und erleichtern die interoperable Nutzung eben dieser. Die Struktur von XMP ist hierarchisch aufgebaut. Eine XMP-Datei (eingebettet oder separat gespeichert) kann mehrere Schemata enthalten. Ein Schema ist eine Sammlung thematisch zusam-

[2]Der NewsCode ist frei zugänglich unter `https://cv.iptc.org/newscodes/`.

mengehöriger Metadaten-Einträge[3]. *Namespaces*[4] stellen sicher, dass Einträge global eindeutig adressiert werden können, auch wenn es ein und den selben Eintrag, zum Beispiel `author`, in mehreren Schemata gibt. Im Gegensatz zu den *controlled vocabularies* der IPTC werden die Werte in anderen Schemata nicht durch den Standard vorgegeben. Die Handhabung etwa in Adobe Photoshop (PS) macht die verschiedenen Schemata teilweise sichtbar, ordnet aber zum Beispiel Einträge des Dublin Cores[5] in die IPTC-Spalte ein. Gibt man in Photoshop, von Adobe Inc. (PS) im Fenster „Dateiinformationen“ unter IPTC-Contact `Creator=Tommy Langstrumpf` ein, so landet der Wert im Eintrag `XMP-dc:Creator` (vgl. `V1-md-edit3.jpg` und Abb. 2.1 S. 9).

2.2 C2PA-Spezifikation

Mit dem vermehrten Aufkommen KI-generierter Bilder und dessen Verbreitung über Online-Plattformen schlossen sich Nachrichten-, Software- und KIUnternehmen mit dem Ziel zusammen, die Vertrauenswürdigkeit der Medienunternehmen zu schützen. Gerade Zeitungsverlage sehen durch den vermehrten Einsatz von KI ihre Glaubwürdigkeit in Gefahr. Die Coalition for Content Provenance and Authenticity (C2PA) schlägt als Lösung kryptografisch gesicherte Metadaten vor, die automatisiert ausgelesen werden sollen. Journalist:innen und Rezipient:innen sollen so bei der Beurteilung der Authentizität eines Bildes unterstützt werden, und KI-generierte Bilder als solche erkennen.

2.2.1 Die Content Authenticity Initiative und ihr Ziel

Die Content Authenticity Initiative (CAI) wurde 2019 von Adobe, der New York Times (NYT) und Twitter (heute X) gegründet. Sie ist nicht der erste Zusammenschluss ihrer Art. Das Project Origin wurde von der British Broadcasting Corporation (BBC), Canadian Broadcasting Corporation (CBC)/Radio Canada , der NYT und Microsoft bereits

[3] Im XMP-Standard wird der Begriff *properties* verwendet.
[4] Der Begriff *Namespace* wird in XMP anders verwendet, als in IPTC. In XMP sind damit die verschiedenen Schemata gemeint.
[5] Der Dublin Core ist ein sehr früher Metadatenstandard, der ursprünglich für Textdokumente konzipiert, und später auf andere Medien ausgeweitet wurde.

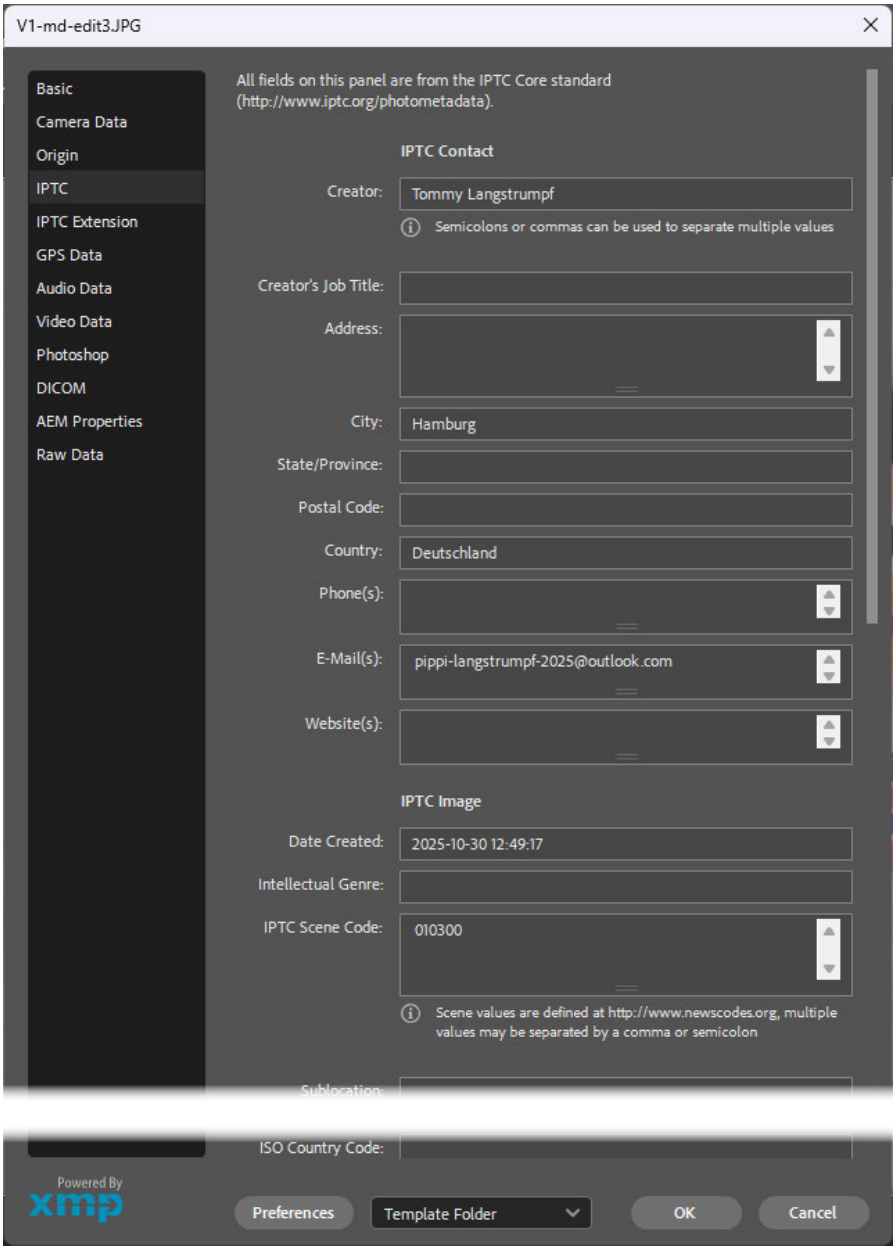


Abbildung 2.1: Fenster „Dateiinformationen“ in PS; geöffnet ist das Bild `V1-md-edit3.jpg`. Das `Date Created` stammt aus der Bearbeitung mit Windows Explorer, der `IPTC Scene Code` wird nur in PS angezeigt.

2018 mit dem Slogan „Protecting Trusted Media“ gegründet (*Project Origin* 2025). Zwei Jahre später entschieden sich alle Akteure, der C2PA beizutreten, um an der Erstellung eines Standardkatalogs mitzuwirken (vgl. Ellis 2022). Auch Arm, Intel und Truepic haben sich dem Ziel angeschlossen[6].
Aus Deutschland sind unter anderem bereits die Deutsche Presseagentur (dpa), die Bundesdruckerei Gruppe GmbH[7], der Axel Springer Verlag, TrustNXT und der Westdeutsche Rundfunk dabei. WDR Mitarbeiter Martin Grohme und Kenneth Warmutz haben zusammen mit Ingo Daniels von der Deutschen Welle eine erste Implementierung im Videostreaming vorgenommen und auf dem Production Technology Seminar der European Broadcast Union (EBU) Anfang 2025 einen Vortrag dazu gehalten (*C2PA Implementation Strategies* 2025).
Die Differenzierung zwischen CAI und C2PA ist nicht leicht, weil es große personelle Überschneidungen gibt. Während die CAI die Interessen nach außen vertritt und sich ihr viele Profiteure auf Seiten der Anwender:innen angeschlossen haben, kümmern sich die C2PA-Mitglieder um die konkrete Ausarbeitung der Spezifikation und ihrer Implementationen.

Die C2PA erarbeitet eine Spezifikation, die Fakten mit einer digitalen Signatur kombiniert[8] und somit ermöglichen soll, die Herkunft medialer Inhalte nachzuvollziehen. Sie beschreibt das Ziel mit dem Wort *provenance*, die CAI bezieht es mit dem Begriff *news provenance* explizit auf den Bereich der Nachrichtenwelt. In der Spezifikation heißt es dazu:

> „Provenance empowers content creators and editors (...) to disclose information about how an asset was created, how it was changed and what was changed. (...) In this way, content with provenance provides indicators of authenticity (...).“ (C2PA V2.1, S. 2)

Damit soll eine Vertrauenskette[9] von der Quelle bis zur Rezipient:in etabliert werden, die als Hinweis auf Authentizität verstanden werden kann.

[6] Eine aktuelle Auflistung aller Mitglieder ist auf der Internetseite `https://c2pa.org/membership/` zu finden.
[7] Die zugehörige Firma D-Trust steuert als Vertrauensdienstleister Zertifikate für C2PA-fähige Softwareanbieter bei.
[8] Im Original „combining statements of fact with a digital signature“ (Temmermans u. a. 2024, S. 2)
[9] Im Original „chain of trust“ (*Project Origin* 2025)

Im deutschsprachigen Raum gibt es nur wenige Internetquellen, die sich einer Übersetzung[10] und damit Interpretation des Begriffs annehmen. Es wird sehr pragmatisch direkt von *Content Credentials* gesprochen oder mit „Echtheitsnachweis“ (Bundesdruckerei GmbH 2024) oder „Inhaltsurhebernachweis“ (Bundesdruckerei GmbH 2025) übersetzt. Content Credentials (CCr)[11] sind das Ergebnis, dass durch automatisiertes Auslesen der C2PA-Metadaten in Form eines Icons sichtbar gemacht werden soll (siehe Abb. 2.2). Die regelmäßige Nennung der Begriffe *provenance* und *authenticity* in direktem Zusammenhang erzeugt beim Lesen der Spezifikation den Eindruck, ein Inhaltsurhebernachweis bedeute automatisch, dass das Bild authentisch ist. Außerdem bleibt der Begriff Authentizität etwas vage. Daher erfolgt eine gesonderte Betrachtung des Konstrukts in Abschnitt 5.1.

Der Inhaltsurhebernachweis soll nicht nur Redaktionen und Verlagen zu Gute kommen. Auch von Rezipient:innen[12] sollen CCr verstanden und benutzt werden, um die Authentizität eines Bildes zu beurteilen. Kreative sollen die CCr nutzen, um ihre digitale Kunst gewissermaßen zu signieren; das würde effizientere Verwertungsstrategien ermöglichen. Die Nutzung von Citizen Media ist für Verlage herausfordernd, weil sich die Verifikation bei unbekannter Quelle oft schwierig gestaltet; mit CCr eben diese Inhalte zuverlässiger und schneller authentifiziert werden, so die Hoffnung.

Abbildung 2.2: Das CCr-Icon

2.2.2 Funktionsweise

Abbildung 2.3 zeigt alle Elemente, die einer Bilddatei durch die C2PA-Spezifikation hinzugefügt werden. Die Zusammenhänge erklären sich am besten an einem Beispiel. Zum Teil konnten gute Übersetzungen für die verwendeten Be-

[10] Die direkte Übersetzung von *provenance* ist laut PONS Wörterbuch „Herkunft“.

[11] Die CAI kürzt Content Credentials mit „Cr“ ab, allerdings wird dieses Akronym auch in anderen Kontexten verwendet, sodass sich hier zwecks klarer Differenzierung für „CCr“ entschieden wurde.

[12] Die Spezifikation benutzt den Begriff *consumer*, was den marktorientierten Hintergrund der beteiligten Instanzen hervorhebt.

griffe gefunden werden, zum Teil sind die englischen Begriffe in diesem Zusammenhang eindeutiger zu verstehen. Angenommen die Fotografin Jane benutzt eine C2PA-fähige Kamera und macht ein Foto. Sie öffnet das Foto in Photoshop und bearbeitet es: sie verändert den Kontrast, den Bildausschnitt und den Farbton. Danach fügt sie IPTC-Metadaten hinzu und exportiert das Foto als jpg-Datei und lässt es einem Zeitungsverlag zukommen.

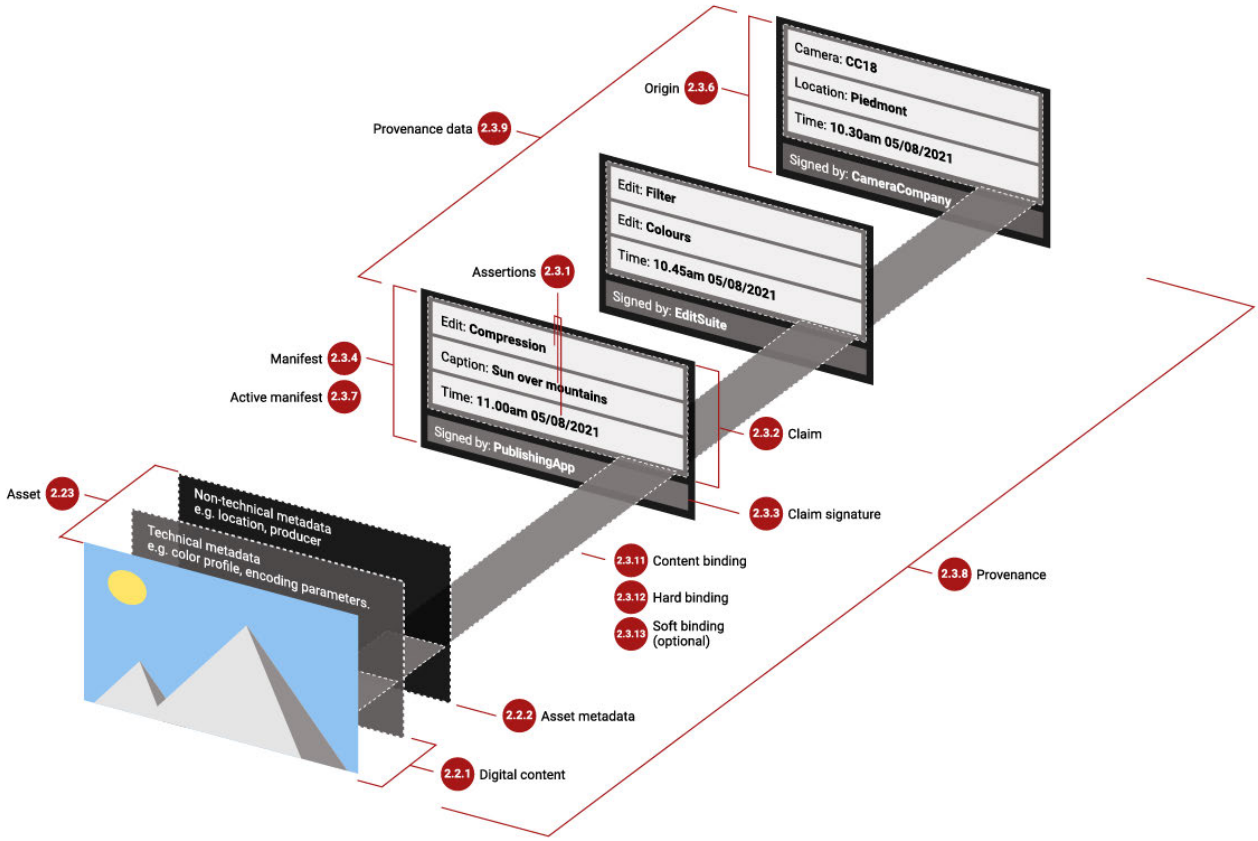


Abbildung 2.3: Funktionselemente der C2PA-Spezifikation, die zusammen *Provenance* erzeugen sollen

> „Each time someone edits or updates the asset using a tool that supports CAI, it adds a new manifest with the actions taken and the certificate of the tool/site; this becomes the *active manifest*, which then references any prior manifests as ingredients.“ (CAI 2025a)

Die eigentlichen Bilddaten (*image data*) zusammen mit den Exif- und IPTC-Metadaten bilden zusammen das *Asset* (vgl. 2.23 in Abb. 2.3). Die Kamera erzeugt eine Aussage (*Assertion*), in der steht, dass das Foto mit genau diesem Kameramodel, mit genau den Einstellungen erzeugt wurde (2.3.1 und 2.36 in der Abb.). Die Kamera hat eine kryptografische Signatur auf ihrem Chip gespeichert. Mit Hilfe dieser Signatur kann eine Behauptung (*Claim*, 2.3.2 in der Abb.) signiert werden (*Claim Signature*, 2.3.3 in der Abbildung). Ein Claim ist dabei lediglich die Sicherheitsverpackung[13], in der mehrere Aussagen und *Content Binding*-Informationen abgelegt werden. Behauptung und Signatur ergeben zusammen ein Manifest (2.3.4 in Abb.). Pro vorgenommene Aktion[14] kann ein Manifest erstellt werden. Das jeweils aktuellste Manifest wird auch *active manifest* genannt. Alle Manifeste werden in eine übergeordnete Struktur, den Manifest Speicher (*Manifest Store*, als *Provenance data* in der Abb. bezeichnet) abgelegt. Die Gesamtheit aller Daten, Bilddaten, Metadaten und Manifeste sollen dann als Herkunftsnachweis interpretiert werden können.

Content Binding

So wie in Abbildung 2.3 und Abbildung 2.4 dargestellt, werden die Manifeste in die Bilddatei eingeschrieben, sodass die Daten zusammen mit den eigentlichen Bilddaten veröffentlicht werden können (vgl. hierzu Abb. 2.4 links oben mit *content* bezeichnet). Die Spezifikation lässt aber auch ein getrenntes Aufbewahren von Bildinhalt und Manifest zu. Damit beide Elemente sicher einander zugeordnet werden können und das Manifest als dazugehörig verifiziert werden kann, muss eine Verknüpfung erfolgen. Das als *Content Binding* bezeichnete Vorgehen kann auf verschiedene Weise durchgeführt werden.

Eine starke Bindung (*Hard Binding*) ermöglicht die Spezifikation durch kryptografische Hashing-Algorithmen. Hashfunktionen sind kryptografische Algorithmen, die Eingangsdaten beliebiger Länge auf einen Wert fester Länge abbilden. Der entstehende Hashwert „ist eine mathematische Prüfsumme“ (*BSI Glossar* 2025), mit deren Hilfe man Daten als

[13] Im Original *tamper-evident* (C2PA V2.1, S. 2)

[14] Im Original *action*; die Spezifikation legt in Abschnitt 18.12 fest, welche Aktionen vermerkt werden müssen bzw. können; etwa `c2pa.resized` oder `c2pa.drawing` kommen auch in `V4-edit3.jpg` vor (vgl. Abschnitt 3.4).

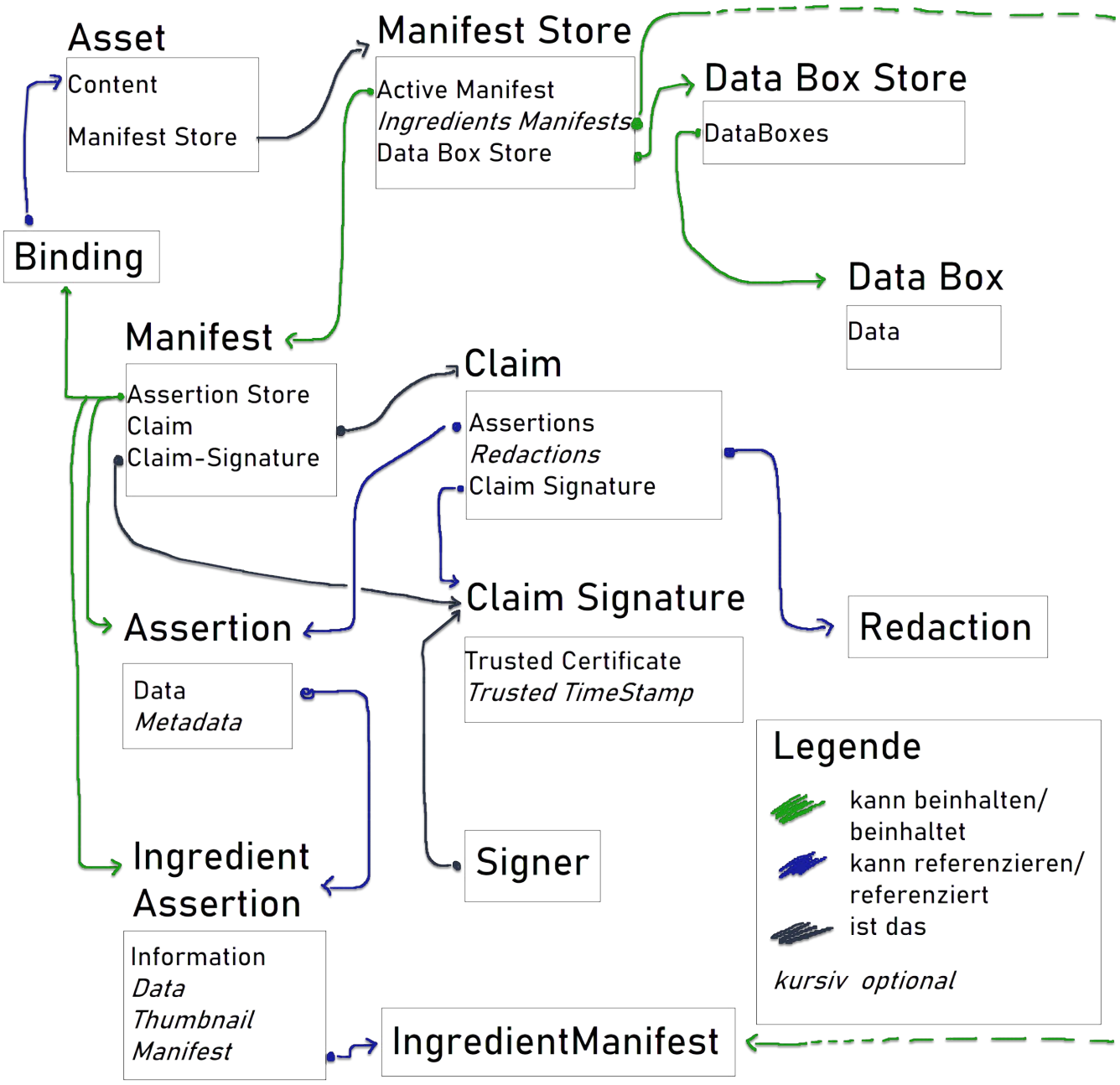


Abbildung 2.4: visuell verbesserte Darstellung des Entitäts-Modell der C2PA aus (C2PA V2.1, S. 55)

unverändert verifizieren, also ihre Integrität feststellen kann. Bilder mit *Hard Bindings* können bearbeitet werden; geschieht dies jedoch mit einer nicht-kompatiblen Software, wird das *Binding* aufgelöst und eine Verknüpfung von Manifest und Asset zwecks Verifikation ist nicht mehr möglich. Bedingt durch die Box-Struktur[15] des Assets können sich die Hashfunktionen auch auf einzelne Boxen beziehen. Während es als „security best practice“ beschrieben wird, wirklich alle Teile einer Datei zu hashen, lässt die Spezifikation[16] es jedoch zu, etwa nur die Bilddaten zu hashen (C2PA V2.1, S. 138).

Für eine schwache Bindung (*Soft Binding*) können Wasserzeichen und Bildsignaturen in die Bildpixel beziehungsweise Metadaten integriert werden (vgl. S. 16). Gerade weil das *Hard Binding* schon durch einfache Änderungen wie Zuschnitt oder Kompression zerbricht, und das Manifest so unbrauchbar wird, ist eine *Soft Binding* Methode als komplementärer Ansatz sinnvoll. Seit Mitte 2024 sprechen Adobe-Mitarbeitende von *durable Content Credentials*, wenn diese nicht nur signierte Hashwerte, sondern auch ein Wasserzeichen enthalten und eine Bildsignatur generiert wird (vgl. Parsons 2024b). So sollen auch Bilder überprüfbar sein, deren CCr mutwillig gelöscht oder aus Versehen beschädigt wurden.

Die Informatikgemeinschaft hat viele Ansätze entwickelt, um KI-generierte Bilder mittels Wasserzeichen als solche zu kennzeichnen. Diese Methoden können auch auf digitale Fotografien angewandt werden. Masood (2025) gibt einen ausführlichen Überblick über Wasserzeichen-Technologien, auch für Text, Audio und Video. Konkrete Ansätze werden zum Beispiel von Saberi u. a. (2024) und Zhao u. a. (2024) aufgezeigt. Die meisten Ansätze erzeugen unsichtbare[17] Wasserzeichen. Werden diese vom KI-Model während des Generierens in das Bild eingebettet, spricht man von *in model* Methoden (vgl. Masood 2025). Diese sind im Allgemeinen robuster gegenüber Manipulation und Zerstörung. Soll ein Bild nachträglich mit einem Wasserzeichen versehen werden, muss ein *post hoc* Verfahren angewandt werden. Zu beiden Zeitpunkten kann das Wasserzeichen statistisch oder kryptografisch gestützt sein, und auf Bit-Ebene oder im Frequenzbereich appliziert werden. Semantische Wasserzeichen (vgl. Zhao u. a. 2024) stellen einen besonderen Ansatz dar. Beitragende aus verschiedenen Disziplinen arbeiten hierbei zusammen. „This

[15] Die Spezifikation nutzt das Datenformat Concise Binary Object Representation (CBOR), um Daten zu strukturieren.
[16] Vergleiche hierzu das Entitätsmodell in Abb. 2.4; Metadaten sind kursiv gesetzt, also optional.
[17] Unsichtbar meint hier mit dem menschlichen Auge nicht wahrnehmbar.

blend of ML [machine learning, Anm. d. Autorin] and cryptography expertise is pushing watermark designs that are both secure and practical“ (Masood 2025). Als Beispiel sei das Tree-Ring-Watermark angeführt, dass *in model* das initiale Rauschen mit einem geheimen Muster versieht. Das beeinflusst den Generierungsprozess und lässt sich durch den invertierten Generierungsprozess nachweisen (vgl. Wen u. a. 2023). Nur wer das geheime Muster kennt, kann das Wasserzeichen finden. Außerdem ist es weder auf Pixelebene noch im Frequenzbereich detektierbar, was es besonders robust gegenüber Manipulation sein lässt.

Für den Fall, dass das Wasserzeichen zerstört wird, soll die Verknüpfung zwischen Datensatz und Bild durch *fingerprinting* ermöglicht werden. Dabei sind nicht die Sensor-Fingerabdrücke gemeint (vgl. Abschnitt 4.2), weshalb sich im Deutschen besser von Bildsignaturen sprechen lässt. Durch maschinelles Lernen erzeugte *perceptual hashes*, werden durch Merkmalsextraktion erstellt. Es handelt sich entsprechend nicht um kryptografische Hashwerte, sondern auf den Bildelementen basierende Hashwerte. „These moreforgiving techniques will equate two images if they are closely similar, unlike cryptographic hashes which require every pixel value to match“ (Collomosse u. a. 2024a, S. 2). Die Bildsignatur wird als *perceptual hash value* in einer Datenbank gespeichert. Sind sich bei einer Bilderrückwärtssuche zwei verglichene Bilder sehr ähnlich, so sind auch ihre Bildsignaturen ähnlich. So kann etwa ein Bild, deren CCr gelöscht wurden, auf das Originalbild mit CCr zurückgeführt werden, oder aber ein Manifest dem Bild zugeordnet werden, auch wenn das *Hard Binding* aufgelöst wurde.
Die Art des *Soft Bindings* muss in einer Assertion festgehalten werden. Es sind nur solche Algorithmen erlaubt, die zuvor in die `softBinding-algorithm-list` [18] eingetragen wurden. Bislang (Stand: 17.03.2025) finden sich in der Liste Wasserzeichen von Digimarc, ATSC, Adobe, Steg.AI, Overlai, Kinetiq (Teletrax), castLabs, Imatag, Nagra NexGuard, und zwei Bildsignatur-Algorithmen von Adobe und aus dem ISO Standard ISO 24138.

Ob stark oder schwach, die Verknüpfung zwischen Asset und Manifest ist elementar. Durch die Nutzung von Signaturen und Hashes wird of von kryptografisch gesicherten Metadaten gesprochen, wenngleich sich das nur auf die Verbindung bezieht, nicht aber

[18] Dazu muss ein Pull Request auf der Seite `https://github.com/c2pa-org/softBinding-algorithm-list` erstellt werden.

auf die eigentlich gespeicherten Informationen. Mit der C2PA-Spezifikation werden viele neue Begriffe, Regeln und Informationsblöcke eingeführt, die helfen sollen, Metadaten so wie von den Urheber:innen gewollt, an den medialen Inhalt zu binden. Darüber hinaus soll auch das Wiederherstellen von Metadaten durch die Auslagerung dieser und eine entsprechende Verknüpfung zwischen Asset und Manifest ermöglicht werden.

3 Versuche

Im Laufe der Literaturrecherche und durch Vorversuche sind einige Behauptungen aufgekommen, die nicht mit Literaturquellen allein überprüft werden konnten. Durch die Versuche sollen entsprechend aufgestellte Hypothesen belegt bzw. widerlegt werden, um im weiteren Verlauf dieser Arbeit die Forschungsfrage zu beantworten.

Die Hypothesen sind von 1 bis 7 durchnummeriert. Die Versuche sind ebenso durchnummeriert, korrespondieren aber nicht zwingend mit den Hypothesen-Ziffern. Für eine Zuordnung siehe Tabelle 3.1. Die verwendete Software ist im Anhang aufgezählt und erläutert (S. 82). Dateien sind überwiegend nach folgendem Schema benannt: `Vx-edit-y.jpg`. Das `x` ordnet die Datei einem Versuch zu, sind in einem Versuch mehrere Arbeitsschritte vorhanden, deutet das `y` an, um welchen Schritt es sich handelt. Auf dem zugehörigen USB-Stick im Ordner `Versuche` können die Protokolle (pdf) nachgelesen werden. Abb. 3.2 zeigt den Inhalt eines Versuchsordners; die Ausgaben von Exiftool sind als `Versuch-x-md.ods` gespeichert; Ausgaben des c2patools wurden als `Vx-edit-y-c2patool-out.txt` gespeichert.
Einer Hypothese können mehrere Versuche zugeordnet sein. Die Auswertung geschieht hypothesenbezogen, im Anschluss werden zusätzliche Erkenntnisse erläutert. Es empfiehlt sicht, das Excel-Dokument `Versuchsauswertung.ods` zu öffnen, um den Ausführungen folgen zu können; die erste Zeile enthält stets den Dateinamen, sodass dort die Metadaten der untersuchten Bilder nachvollzogen werden können. Die relevantesten Bilder aus den Versuchen sind in Abbildung 3.1 zu sehen.

Tabelle 3.1: Hypothesen und Versuchszuordnung

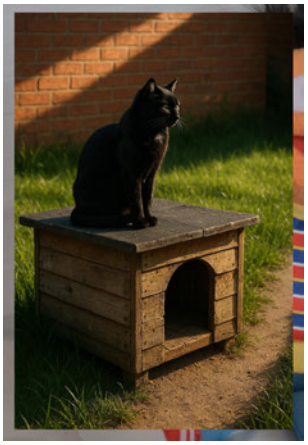
Nr.	Hypothese	Versuche
1	Metadaten können gelöscht und bearbeitet werden.	
1-A	Eine Fotografie enthält EXIF-Metadaten. Diese können gelöscht oder geändert werden. Dabei kann entweder der Wert eines Feldes gelöscht oder geändert werden oder der gesamte Eintrag.	1, 4
1-B	Die IPTC-Metadaten einer Fotografie lassen sich mit Hilfe von Adobe Photoshop und GIMP bearbeiten und löschen.	1, 4
1-C	Die C2PA-Metadaten eines KI-generierten Bildes lassen sich löschen, sodass keine CCr mehr verifiziert werden können.	6
2	Social Media Plattformen verändern Metadaten von auf ihnen hochgeladenen Bildern.	
2-A	Facebook verändert einzelne Metadaten-Einträge.	5
2-B	Instagramm verändert einzelne Metadaten-Einträge.	5
2-C	Mastodon verändert einzelne Metadaten-Einträge.	5
2-D	X verändert einzelne Metadaten-Einträge.	5
3	KI-Bilder enthalten CCr	
3-A	Durch Firefly erzeugte Bilder enthalten C2PA-Manifeste und können mit Hilfe der CAI-Verify-Seite verifiziert werden.	6
3-C	Die mittels CoPilot erzeugten Bilder enthalten C2PA-Manifeste und können mit Hilfe CAI-Verify-Seite verifiziert werden.	8
3-D	Die mittels Dall-E erzeugten Bilder enthalten C2PA-Manifeste und können mit Hilfe der CAI-Verify-Seite verifiziert werden.	9
4	CCr speichern Bearbeitungsschritte und Verify-Seiten zeigen das.	
4-A	C2PA-Metadaten zeigen, dass ein Bild mehrmals geöffnet, bearbeitet und dann exportiert wurde.	4
4-B	Die Verify-Seite zeigt lediglich Informationen aus dem aktiven Manifest an, sodass nur der letzte Bearbeitungsschritt ersichtlich ist.	4
5	Das C2PA-Feld `CBOR:Title` lässt sich nicht mit einfachen Mitteln verändern.	6
6	Es ist innerhalb kürzester Zeit möglich, die Metadaten eines Kamerabildes auf ein KI-generiertes Bild zu übertragen, sodass die Analyse der Metadaten keinen Rückschluss auf den künstlichen Ursprung des KI-generierten Bildes zulässt.	3
7	Eine manipulierte digitale Fotografie kann mit validen CCr versehen werden.	4

(a) Ausgangsbild `V4.jpg`, aufgenommen mit einer Smartphone-Kamera

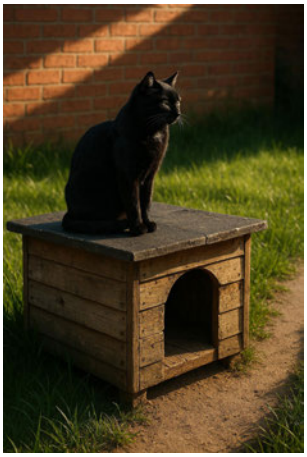
(b) `V4-edit1.jpg`; mit GIMP bearbeitet

(c) `V4-edit3.jpg`; mit Photoshop und Firefly bearbeitet.

(d) `V5-cai.jpg`; Beispielbild der CAI, verwendet in Versuch 5.



(e) `V3-fake.jpg`; Bild mit zwei Ebenen, unten ein Kamerabild, oben ein synthetisches Bild



(f) `V3-cat-with-CCr.jpg`; nach Zuschnitt kein visueller Rückschluss auf synthetischen Ursprung möglich

Abbildung 3.1: Bilder aus den Versuchen

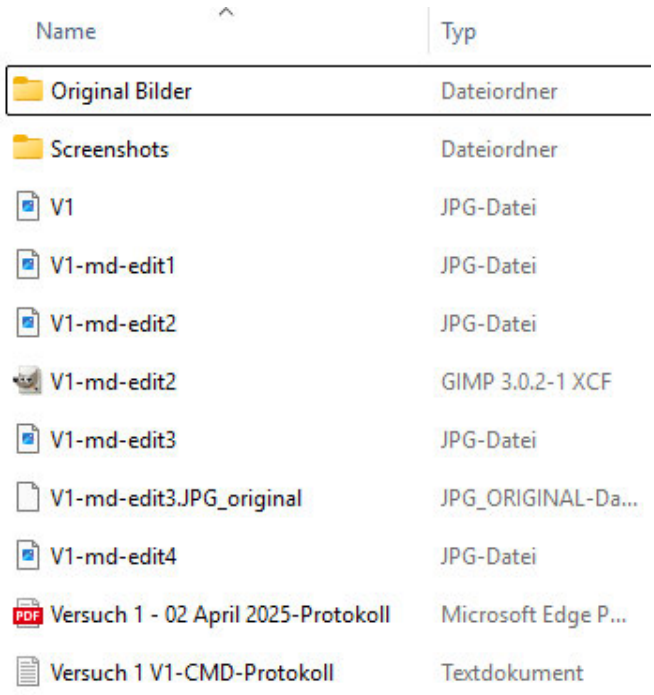


Abbildung 3.2: Inhalte eines Versuchsordners (Versuch 1)

3.1 H1 - Metadaten können gelöscht und bearbeitet werden

In Versuch 1 und 4 wurden Metadaten mit einfachen Mitteln bearbeitet oder gelöscht. Mit dem Windows Explorer lassen sich nur bestimmte Exif-Einträge bearbeiten, unberührt bleiben hingegen die kameraspezifischen Einträge, hier die `Nikon` und `NikonCustom` Einträge in `V1-md-edit1.jpg`. Ausnahme bildet der Eintrag für die Seriennummer `Nikon:SerialNumber` in `V1.jpg`, welcher durch die Bearbeitung gelöscht wurde; der manuell eingegebene Wert wird im Feld `ExifIFD:SerialNumber` vermerkt (`V1-md-edit1.jpg`). Durch `XMP-dc` und `XMP-microsoft` Einträge wird klar, dass Microsoft-Software zur Bearbeitung verwendet wurde.
Ähnlich ist es bei der Bearbeitung mit GNU Image Manipulation Program (GIMP). Die Software legt ein ICC-Farbprofil an und speichert Informationen darüber in den Metadaten (`V1-md-edit2.jpg`). Gleichzeitig löscht GIMP alle kameraspezifischen Einträge der Nikon, wenn Exif-Daten beim Export nicht ausgewählt werden und erstellt

`XMP-xmpMM` sowie `IFD0` Einträge, welche manuell hinzugefügte Werte enthalten und zeigen, welche Bearbeitungsschritte vorgenommen wurden (etwa `XMP-dc:Creator` und `XMP-xmpMM:HistoryChanged`). Hieraus wird ersichtlich, dass nur die Metadaten bearbeitet wurden! Im Vergleich mit anderen Versuchen fällt außerdem auf, dass die `IFD1`-Einträge zum Miniaturbild (*thumbnail*) nicht gelöscht werden, wenn nur Metadaten bearbeitet wurden.
Mit dem Kommandozeilenwerkzeug Exiftool hat man die weitreichendste Kontrolle[1] über Metadaten. Es lassen sich fast alle untersuchten Einträge bearbeiten, bis auf das C2PA-spezifische Schema `CBOR`. Es lassen sich alle Metadaten löschen, die nicht für die korrekte Darstellung der Bilddatei notwendig sind. Auch Photoshop löscht kameraspezifische Metadaten und somit Informationen über den Ursprung des Bildes (`V4-edit2.jpg`). Allerdings variiert das Verhalten der Software abhängig von den Exporteinstellungen.
Möchte man ein Bild zurückdatieren, um den Eindruck zu erwecken, es sei aus einer bestimmten Zeit, müssen viele Einträge bearbeitet werden. Das ist nicht unmöglich, aber schon ein vergessener Wert führt zu Inkonsistenz und dementsprechend zu Skepsis bei der Analyse der Metadaten. Außerdem müssen die Felder `XMP-xmpMM:HistoryChange` und `XMP-xmpMM:HistoryWhen` gelöscht werden, um keinen Hinweis auf die Bearbeitung zu hinterlassen.

Die Hypothese lässt sich also bestätigen, allerdings können C2PA-spezifische Metadaten nicht mit einfachen Mitteln bearbeitet (aber gelöscht) werden. Die konsistente Fälschung von Datums-Angaben ist aufwendig.

[1] Eine stetig aktualisierte Liste aller mit Exiftool bearbeitbaren Metadaten findet sich auf der Internetseite `https://exiftool.org/TagNames/index.html`.

3.2 H2 - Social-Media-Plattformen verändern Metadaten

In Versuch 5 wurden drei Bilder auf folgende Plattformen hochgeladen: Facebook, Instagram, X und Mastodon[2].
Das Bild[3] `V5-cai.jpg` enthielt im Eintrag `XMP-dcterms:Provenance` einen Link in die Adobe-Cloud, sodass das c2patool das Manifest auslesen und validieren konnte. Das KI-generierte Bild `V5-OpenAI.png` hat die C2PA-Daten direkt in die Datei eingebettet. Das Bild `V5-plant.jpg` enthielt keine CCr.
Auf allen Plattformen wurden die `CBOR` Einträge gelöscht, sodass anschließend keine Validierung mit der Verify-Seite stattfinden konnte. Auch `Exif`, `XMP`, `IPTC`, `Composite`, `IFD0` und `Adobe` -Metadaten wurden gelöscht. Zusammengefasst alles, was Rückschlüsse auf den Ursprung der Datei zulassen könnte. Facebook und Instagram verwenden augenscheinlich den gleichen Algorithmus, um den Dateinamen zu ändern (vgl. etwa `fb-download-cai[...].jpg` und `insta-download-cai[...].jpg`). Dass im Dateinamen und im Eintrag `IPTC:SpecialInstructions` eine Art Nutzer:innen-ID und/oder Bild-ID vermerkt wird, um die Verbreitung des Bildes auf den Plattformen zu verfolgen, liegt nahe. Auch X und Mastodon verändern den Dateinamen (`X-download-cai[...].jpg` und `masto-download-cai[...].jpg`), ein Muster konnte jedoch in diesem Versuch nicht erkannt werden.

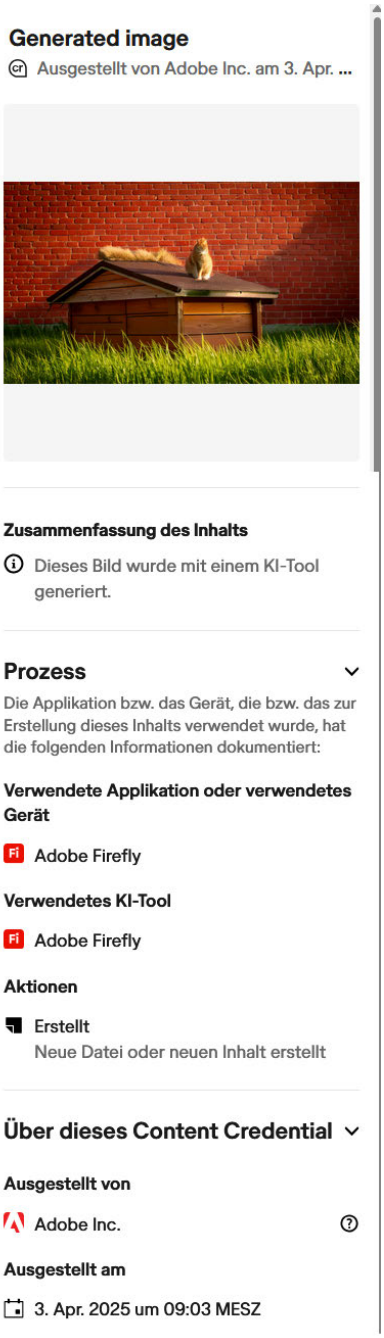
Sobald ein Bild auf eine der untersuchten Social-Media-Plattformen hochgeladen wurde, sind die Metadaten, die Auskunft über den Ursprung des Bildes geben könnten, verloren. Vielleicht ist Meta in der Lage, den:die ursprüngliche:n Nutzer:in zu identifizieren, der:die ein Bild zuerst hochgeladen hat, aber ob diese Informationen mit anderen Instanzen geteilt oder offengelegt werden, ist unbekannt. Somit ist die Hypothese bestätigt und das Ausmaß der Löschung von Metadaten durch Social-Media-Plattformen dargelegt. Zwar behaupten die Plattform-Betreiber:innen die C2PA-Spezifikation bereits implementiert zu haben, ein CCr-Pin konnte bei den Posts jedoch nicht gefunden werden.

[2] Es sei angemerkt, dass verschiedene Mastodon-Server verschiedenes Verhalten aufzeigen können. Die Plattform wurde als nicht-proprietäre Alternative mit in den Versuch aufgenommen.
[3] Die Bilddatei ist der Internetseite `https://contentcredentials.org/` entnommen.

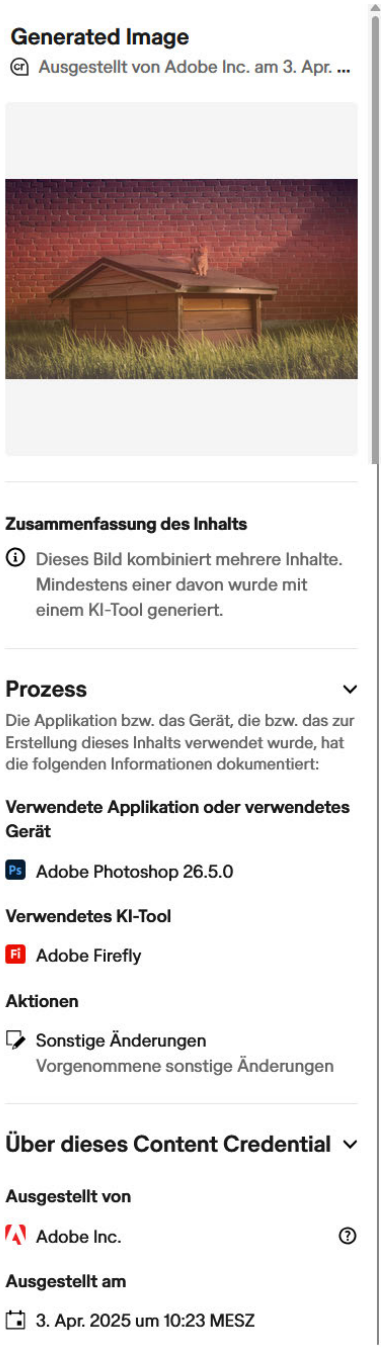
3.3 H3 - KI-generierte Bilder enthalten Content Credentials

In der Versuchen 6, 8 und 9 wurde generative künstliche Intelligenz von Adobe, OpenAI und Microsoft untersucht. Alle untersuchten KIs binden CCr in ihre Bilder ein.
Auch durch die Nutzung der Funktion *generative fill*[4] innerhalb Photoshops werden CCr erzeugt. Eine Validierung mit der Verify-Seite erzeugt die Aussage „Dieses Bild wurde mit einem KI-Tool generiert“ wie in Abbildung 3.3a zu sehen ist. Als Anwender:in hat man bei dieser Form der Bearbeitung nur begrenzt Einfluss auf die in das Manifest inkludierten Informationen. Als Aktion wird entweder „erstellt“ oder „sonstige Bearbeitungen“ (Abb. 3.3) in die Assertion geschrieben (vgl. *Content Credentials im Überblick* 2025). Wurden zuvor in Photoshop persönliche Daten hinzugefügt oder Social-Media-Konten verknüpft, so wird das auch auf der Verify-Seite angezeigt.
Der chatBot von OpenAI, chatGPT erstellt mittels Prompt eine jpg-Datei (`V9.jpg`), diese enthält im Gegensatz zur png-Datei (`V9-png.png`) keine CCr.
In Microsoft Designer erzeugte Bilder enthalten nur dann CCr, wenn zuerst das Dateiformat ausgewählt, dann die Einbindung von CCr aktiviert, und anschließend erst auf Download geklickt wurde. Der Prozess ist fehleranfällig und das Einbinden von CCr zum Zeitpunkt der Versuche optional. Die von Microsoft Designer exportierten Bilder `V8-copilot.png` und `V8-msD-edit1-jpgV2.jpg` erzeugen einen Fehler auf der Verify-Seite, das c2patool vermerkt unter `ingredientsAssertionURI` den Fehler `assertion.required.missing`. Das Manifest beinhaltet nur Hashes und Signaturen. Die ausgelesenen Metadaten verraten aber, welcher Name[5] in dem Microsoft-Konto angegeben wurde, und es ist vermerkt, dass es sich um eine kreative Arbeit[6] handelt. Der `CBOR:ActionsSoftwareAgent` wird als `Image Creator from Designer` identifizierbar und die `CBOR:ActionsDescription` verrät eindeutig, dass das Bild KI-generiert ist. Exiftool vermerkt einen Fehler:
`Incorrect JUMBF sequence numbering (should) start from 0, not 1`, was

[4] Mit *generative fill* hat man die Möglichkeit, einen zuvor markierten Bereich des Bildes mit Hilfe einer Prompteingabe zu füllen.
[5] In `V8-msD-edit1-png.png` wurde als Kontoname „Nina Blumenthal“ gesetzt. Diese Angabe findet sich im Eintrag `JSON:AuthorName` wieder.
[6] Vgl. hierzu `JSON:Type = CreativeWork` in `V8-msD-edit1-jpgV2.jpg`.



(a) Verify-Ergebnis zeigt, dass das Bild `V6-firefly.jpg` mit Adobe Firefly generiert wurde.



(b) Verify-Ergebnis zeigt, dass mindestens ein Teil des Bildes `V6-edit3.jpg` mit Adobe Firefly generiert und zusätzlich aus PS exportiert wurde.

Abbildung 3.3: Verify-Ergebnis aus Versuch 6; rechts die bearbeitete Version mit Hinweis auf „sonstige Bearbeitungen“.

wahrscheinlich der Grund dafür ist, dass weder die Verify-Seite noch das c2patool das Bild validieren können, obwohl C2PA-Metadaten vorhanden sind.

Die derzeitigen Implementationen der C2PA-Spezifikation in generativer KI sind noch fehlerbehaftet und die CCr oft noch optional. Dennoch kann die Hypothese insofern bestätigt werden, als das alle angegebenen Unternehmen bemüht sind, die Technologie zur Verfügung zu stellen.

3.4 H4 - C2PA-Metadaten zeigen, dass ein Bild mehrmals bearbeitet wurde

Der Versuch 4 sollte mehrere Fragen klären. Eine digitale Fotografie wurde sowohl mit GIMP als auch mit PS bearbeitet.
Erst wurden mit GIMP sowohl Metadaten als auch Bilddaten bearbeitet (`V4-edit1.jpg`, wie in Abb. 3.1b), danach wurde die Datei in PS geöffnet, die Funktion „Content Credentials (Beta)“ aktiviert und die Bilddatei direkt ohne Bearbeitung als `V4-edit2.jpg` exportiert. Überprüft man die Datei mit der Verify-Seite, erscheint der Hinweis „Einige Bearbeitungen oder Aktivitäten wurden möglicherweise nicht aufgezeichnet“ (Abb. 3.4). Der Hinweis ist bei jeder digitalen Fotografie zu erwarten, die nicht mit einer C2PA-kompatiblen Kamera erstellt wurde! Weil PS sowohl `Exif`, als auch `IDF0` Einträge löscht, kann weder mittels Verify-Seite noch Exiftool ermittelt werden, dass die Datei zuvor mit GIMP bearbeitet wurde.
Nach der Bearbeitung der Bildebene in PS (`V4-edit3.jpg`, wie in Abb. 3.1c) zeigt die Verify-Seite ausführlichere Informationen über die Bearbeitungsschritte (siehe Abb. 3.5). Nicht verwendete *ingredients* werden ebenso angezeigt. Alle durch Firefly in PS erstellten Varianten und die original Fotografie werden aufgeführt wie in Abbildung 3.6 zu sehen ist. Keines davon enthält CCr, das Resultat jedoch bekommt das kleine Icon und Adobe bestätigt, dass Pippi Langstrumpf dieses Bild erstellt, und dazu Photoshop und Firefly verwendet hat. Die dargestellten Informationen über die Bearbeitungsschritte sind nicht falsch, lassen aber auch keine eindeutige Aussage zu. Welches der Bestandteile wurde mit Firefly generiert? Wann wurden die Bestandteile erstellt? Was sind unbekannte Änderungen? Wer verbirgt sich hinter Pippi Langstrumpf?

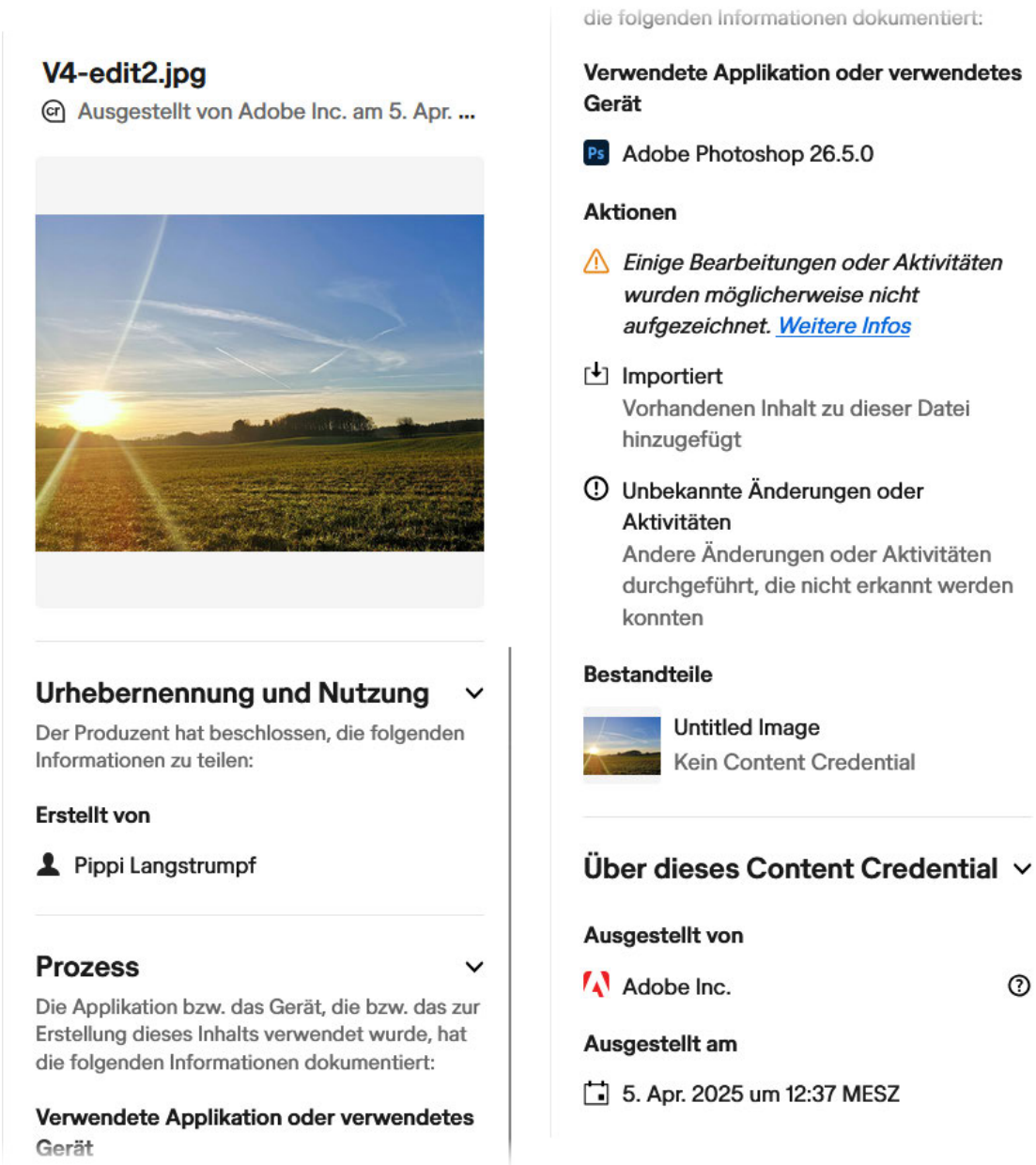


Abbildung 3.4: Verify-Ergebnis zu `V4-edit2.jpg`. Das Bild wurde zuvor in GIMP bearbeitet und enthielt durch Öffnen und Exportieren in Photoshop gültige CCr.

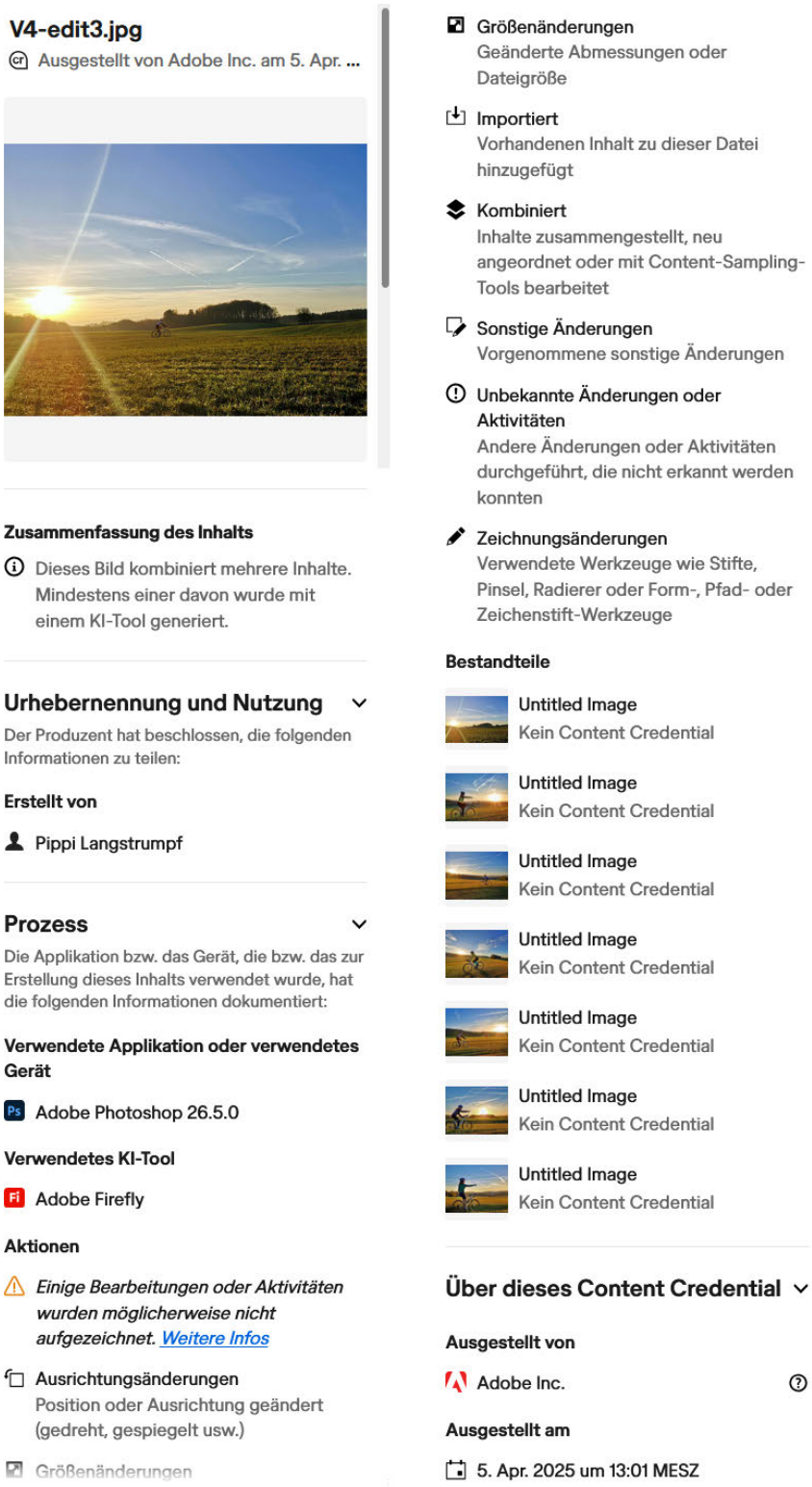


Abbildung 3.5: Verify-Ergebnis zu `V4-edit3.jpg`. Das Bild wurde mit PS *generative fill*, also mit der KI Firefly bearbeitet. Es enthält gültige CCr.

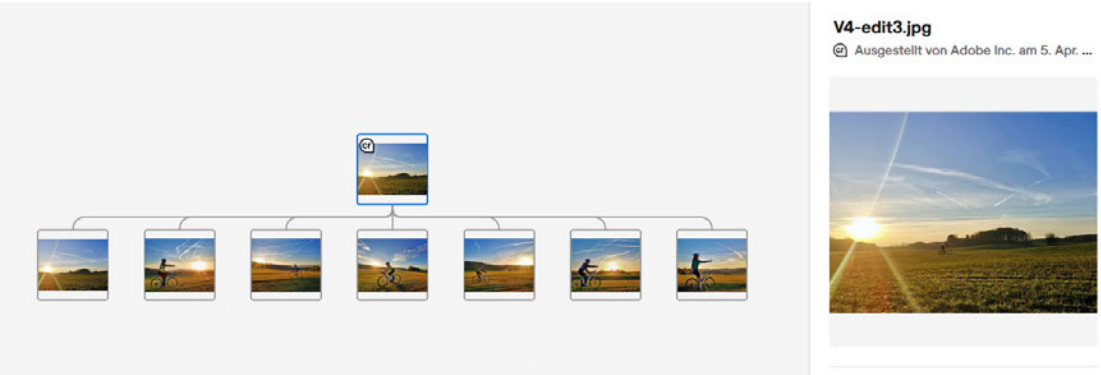


Abbildung 3.6: Visuelle Darstellung der *ingredients* die bei der Erstellung von `V4-edit3.jpg` in PS durch *generative fill* mit Adobe Firefly erzeugt wurden. Darstellung ist der Verify-Seite entnommen.

Der Einbettung der Daten zu Änderungen und Aktivitäten kann in PS widersprochen werden (vgl.Abb. 3.9 auf Seite 36). Was Rezipient:innen auf der Verify-Seite zu sehen bekommen, ist dementsprechend stark abhängig vom Willen und der Ehrlichkeit des:der Urheber:in. Eine Analyse der Metadaten und das Auslesen mittels c2patool verraten hier mehr als die Verify-Seite. Es ist erkennbar, dass die Datei `V4-edit3.jpg` sieben Ebenen enthält (u.a. durch `CBOR:ActionsDigitalSourceType`), die Reihenfolge und genaue Bezeichnung der Bearbeitungsschritte lässt sich am Wert des Eintrags `CBOR:ActionsAction` und der im Manifest vermerkten `c2pa.actions` nachvollziehen. Die `credentialSubectId` unter `JSON` verrät, dass alle Bestandteile von ein und demselben Adobe-Konto hinzugefügt wurden.
Weder Exiftool noch c2patool erkennen das ursprüngliche Aufnahmedatum[7] der Fotografie (01.12.2024),und auch nicht das manipulierte[8] Datum (01.01.2024). Es sieht so aus, als sei das Bild am 05.04.2025 von Pippi Langstrumpf erstellt worden.
Die Datei `V4-edit3.jpg` enthält nur ein Manifest, trotz umfangreicher Bearbeitung. Zur Beantwortung des zweiten Teils der Hypothese wurde das synthetische Bild von chatGPT `V9-png.png` verglichen; es enthält zwei Manifeste, obwohl es nicht bearbeitet wurde. Entsprechend hat die Anzahl der Manifeste keine Aussagekraft hinsichtlich der Art oder Intensität von Bearbeitungen.

[7] Das Aufnahmedatum ist in `V4.jpg` nachvollziehbar unter `System:FileModifyDate` .

[8] Wurde mit Exiftool in das Feld `Composite:DateTimeCreated` geschrieben (`V4-edit1.jpg`), aber von PS gelöscht (`V4-edit2.jpg`).

Die C2PA-Metadaten und auch die Verify-Seite der CAI zeigen, dass ein Bild bearbeitet wurde, und in unterschiedlichem Ausmaß sind auch einzelne Bearbeitungsschritte erkennbar. Die Menge an Informationen ist von den Einstellungen in PS abhängig. Die Informationen auf der Verify-Seite der CAI lassen Interpretationsspielraum. So ist etwa die Region, in der Bearbeitungen vorgenommen wurden, nicht ersichtlich. Auch mehrfache Exporte sind nicht ersichtlich.

3.5 H5 - Der C2PA-Eintrag `CBOR:Title` lässt sich nicht mit einfachen Mitteln bearbeiten

Mit Versuch 6 sollte genauer untersucht werden, ob C2PA-Metadaten einfach manipuliert werden können. Der Eintrag `CBOR:Title` wurde beispielhaft gewählt. In `V6-firefly.jpg` etwa wird durch den Wert `Generated Image` der synthetische Ursprung deutlich. Zusammen mit den Einträgen `ActionsDigitalSourceType` und `ActionsSoftwareAgent` lässt sich ein KI-generiertes Bild eindeutig als solches identifizieren (siehe etwa `V9-png.png`).
Im Versuch 1 wurde bereits festgestellt, dass sich C2PA-spezifische Einträge nicht direkt bearbeiten oder löschen lassen. In Versuch 6 wurde erneut mit Exiftool probiert, in `V6-firefly-edit1.jpg` (direkte Kopie des Originals `V6-firefly.jpg`) das Feld `CBOR:Title` zu bearbeiten. Es lässt sich nicht bearbeiten. Der eingebebene Wert `directly fotographed`[9] wird stattdessen in `XMP-dc:Title` geschrieben und eine Fehlermeldung ausgegeben:
`Warning: Sorry, CBOR:Title doesn't exist or isn't writable [...]`.
Dennoch wird das *Content Binding* (vgl. Abschnitt 2.2.2) dadurch aufgelöst, weil sich der Hashwert auf Bild- und Metadaten bezog. Das c2patool gibt unter `failure` den Eintrag `assertion.dataHash.mismatch` aus und die Verify-Seite zeigt, dass die Datei möglicherweise manipuliert wurde (Abb. 3.7).

Die `CBOR` Einträge lassen sich nicht mit einfachen Mitteln manipulieren. Allerdings ist bei weiteren Untersuchungen im Versuch 6 auch herausgekommen, dass es Schwach-

[9] Der Wert wurde absichtlich falsch geschrieben, um die Manipulation erkennbar zu machen.

stellen gibt, wodurch sich gefälschte Metadaten mit validen CCr versehen lassen, was den Eindruck erwecken könnte, die Metadaten seien authentisch. Außerdem legen Neal Krawetz *Exploits* nahe, dass mit etwas mehr Geduld und Können weitere Schwachstellen gefunden und ausgenutzt werden können (mehr dazu in Abschnitt 5.4).

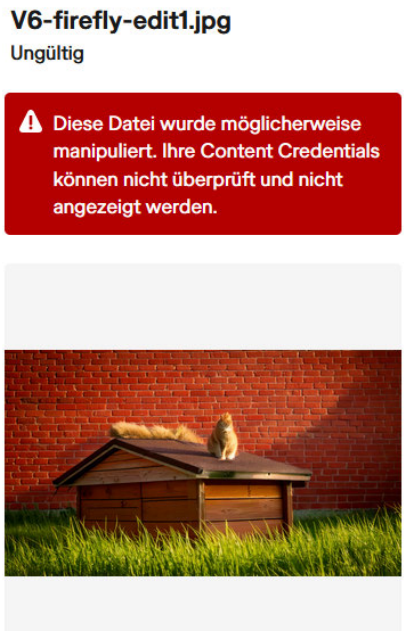


Abbildung 3.7: Verify-Ergebnis von `V6-firefly-edit1.jpg`; durch Bearbeitung der Metadaten wurde das *Content Binding* aufgelöst.

3.6 H6 - Die Metadaten einer digitalen Fotografie können auf ein KI-generiertes Bild übertragen werden

Die ganze Hypothese ist in Tabelle 3.1 nachzuvollziehen. Versuch 3 wurde zuletzt durchgeführt, und verwendet die Ursprungsdateien aus den anderen Versuchen. Ausgangsdateien sind `V3-ernie.jpg` und `V3-landschaft.jpg` sowie eine Kopie der durch chatGPT generierten Katze `V3-OpenAI-cat.jpg`. Der erste Schritt erzeugte die Datei

`V3-ernie-ps-kopieren Kopie.jpg`, welche die gleichen Metadaten wie das Ernie-Bild enthält, obwohl die Landschaft zu sehen ist. Es wurde einfach der Bildinhalt von `V3-landschaft.jpg` mittels Str+C und Str+V auf das Ernie-Bild gelegt und exportiert. Es entstehen die PS-typischen Einträge zum ICC-Farbprofil, die `Nikon` Einträge werden gelöscht, `Exif` und `Composite` Einträge deuten auf die Nikon-Kamera hin, obwohl das gezeigte Bild mit einem Smartphone erstellt wurde!
Im nächsten Schritt wurde das Gleiche mit `V3-OpenAI-cat.jpg` durchgeführt. Das Ergebnis[10] `V3-fake.jpg` enthält keine CCr, weil die jpg-Datei von OpenAI keine CCr enthält (vgl. Abschnitt 3.3). Also wurde im nächsten Schritt das Bild `V3-fake.jpg` in PS importiert, die Funktion „Content Credentials (Beta)“ aktiviert und nach einem Zuschnitt die Datei unter dem Namen `V3-cat-with-CCr.jpg` (vgl. Abb. 3.1f) exportiert. Das Verify-Ergebnis findet sich in Abb. 3.8, es deutet nichts auf einen synthetischen Ursprung hin, Pippi Langstrumpf hat das Foto in Photoshop bearbeitet und wie es aussieht nur einen Zuschnitt vorgenommen. Der Versuch hat inklusive Dokumentation 60 Minuten gedauert. Um die Manipulation noch perfekter aussehen zu lassen, könnten zusätzliche IPTC-Metadaten mit PS hinzugefügt, und weitere künstlerische Anpassungen vorgenommen werden, `Exif`-Einträge mit Exiftool geschrieben, und so eine größere Informationsmenge auf der Verify-Seite erzeugt werden, die vom eigentlich unverifizierten Ursprung ablenkt.

Es erfordert zwar mehrere Schritte, aber prinzipiell ist es möglich, ein KI-generiertes Bild so zu bearbeiten, dass auf der Verify-Seite der Eindruck entsteht, es handele sich um eine bearbeitete digitale Fotografie. Zeloof (2023) hat sogar Manifeste erzeugt, die direkt auf eine Leica M11 Kamera als Ursprung verweisen, wenngleich hier mehr Schritte notwendig waren. Auch können Metadaten von einem Bild auf sehr einfache Weise in ein anderes Bild geschrieben werden; mit Exiftool wäre das sogar noch umfassender gelungen.

Damit ist die Hypothese bestätigt, wenngleich eine:r geübten Forensiker:in die Manipulation durch Untersuchung der Containerebene (siehe Abschnitt 4.3) auffallen könnte.

[10] Das gleiche Ergebnis hätte man mit einem Zwischenschritt mehr auch mit Hilfe der Datei `V9-ApenAi-png.png` erzielen können.

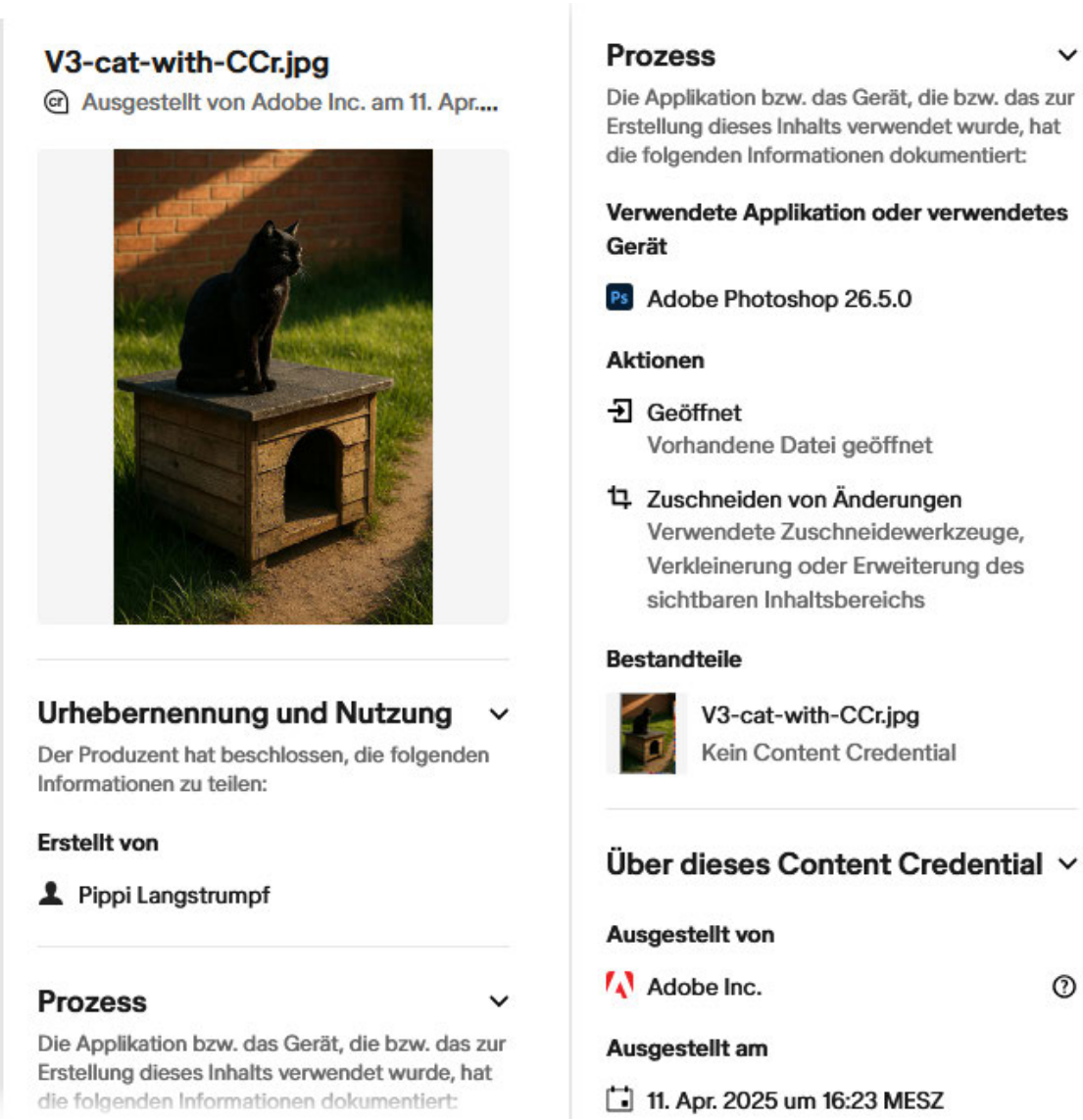


Abbildung 3.8: Verify-Ergebnis zu `V3-cat-with-CCr.jpg`. Die ursprüngliche Datei war `V9.jpg`. Der synthetische Ursprung ist hier nicht erkennbar.

3.7 H7 – Eine manipulierte digitale Fotografie kann mit validen CCr versehen werden

Die Glaubwürdigkeit eines Bildes wird sowohl vom sichtbaren Inhalt (Bilddaten) und dessen Plausibilität, als auch vom schlüssigen Zusammenhang zwischen Bilddaten und Metadaten beeinflusst (vgl. Abschnitt 5.1). Dass sich manipulierte Metadaten mit validen CCr versehen lassen, zeigt Versuch 6.

Die Datei `V6-firefly.jpg` (vgl. Abb. 3.3) wurde im Browser mit Adobe Firefly erstellt. Sie enthält gültige CCr. Mit `V6-firefly-edit1.jpg` wurde Hypothese 5 untersucht. Die Bildebene wurde mit PS bearbeitet, so entstand `V6-firefly-edit2.jpg` mit gültigen CCr, die Verify-Seite zeigt, dass PS benutzt wurde. Die Änderungen an IPTC-Metadaten (`XMP-dc` u. `XMP-iptcCore` Einträge wurden vorgenommen) mit PS in `NV6-edit3.jpg` führen nicht zu einem Bruch des *Content Bindings*. Es können also beliebige Metadaten in die Datei geschrieben werden.
`NV6-edit4.jpg` ist eine Kopie von `NV6-edit3.jpg`, enthielt zunächst also gültige CCr. Mit Exiftool wurde das Feld `XMP-xmpMM:DocumentID` manipuliert. Die Datei kann durch das aufgelöste *Content Binding* nicht validiert werden. Importiert man diese Dati aber in PS und exportiert sie gleich wieder unter neuem Namen `NV6-edit5.jpg`, so enthält diese Datei valide CCr, obwohl die Metadaten gefälscht wurden. Theoretisch hätten auch Bilddaten manipuliert werden können. Wie Neal Krawetz auch mehrmals betont: C2PA lässt falsche Metadaten valide aussehen (vgl. Krawetz 2024c, Krawetz 2024d). Das liegt daran, dass bei jeder Bearbeitung – ob an Metadaten oder Bilddaten vorgenommen – der Hashwert neu berechnet werden muss, das alte Manifest also nicht validiert werden kann, wenn nur der neue Hash zur Verfügung steht. C2PA erneuert das Manifest, ganz gleich, ob die vorgenommenen Änderungen eine Manipulation darstellen oder nicht. Der Versuch beantwortet nicht die eigentliche Hypothese, zeigt aber, dass valide Content Credentials kein Garant für richtige Informationen sind.
In Versuch 4 wurde eine digitale Fotografie mit GIMP manipuliert (`V4-edit1.jpg`, vgl. Abb. 3.1b, Bilddaten wurden geändert). Durch Öffnen und Exportieren in PS erhält die entstehende Bilddatei `V4-edit2.jpg` valide CCr (vgl. Abb. 3.4). Die bereits manipulierte Datei wurde in PS bei aktivierter „Content Credentials (Beta)“-Funktion

weiter bearbeitet und exportiert (`V4-edit3.jpg`, Abb. 3.1c). Die Verify-Seite zeigt viele Informationen (vgl. Abb. 3.5), und die validen CCr lassen vermuten, dass das Bild echt ist, nur eben ein bisschen bearbeitet. Dabei basiert das Bild auf einer manipulierten Bilddatei!

Es ist folglich möglich, eine digitale Fotografie zu manipulieren, und sie durch valide CCr dennoch authentisch wirken zu lassen. Außerdem können Metadaten manipuliert, und durch erneuten Export die Spuren der Manipulation verwischt werden.

3.8 Weitere Erkenntnisse

Durch die Versuche wurden weitere Erkenntnisse erlangt, die Wichtigsten werden hier kurz aufgeführt.

Nimmt man an einem KI-generierten Bild Änderungen in PS vor, so wird aus `trainedAlgorithmicMedia` der Wert `compositeWithTrainedAlgorithmicMedia` (`V6-firefly-edit2.jpg`). Das wirft Fragen auf, weil auch digitale Fotografien die mittels KI bearbeitet wurden, letzteren Wert erhalten (vergleiche hierzu `V4-edit3.jpg`). Der Eintrag lässt keine Aussage zu, ob alle, eine bestimmte Anzahl und wenn ja welche Bestandteile eines Bildes KI-generiert sind. Das Bild als Ganzes muss folglich als KI-generiert angenommen werden, weil keine Differenzierung vorgenommen werden kann.

Die Mehrheit der generativen KIs erzeugt quadratische Bilder und gibt diese standardmäßig als png-Datei aus. Bildformat und Dateiformat können unter Umständen also auf ein KI-generiertes Bild hindeuten. Insgesamt ist zu beobachten, dass KI-generierte Bilder vergleichsweise wenig Metadaten enthalten. Diese Anhaltspunkte können bei der Analyse als Hinweis auf einen synthetischen Ursprung verwendet werden (vgl. Abschnitt 4.3).

Sowohl Adobe als auch Microsoft erlauben es Nutzer:innen, ihren Account-Namen zu ändern. Der Name erscheint als Creator in den C2PA-Metadaten. Dadurch wird der Eintrag `author` in den Metadaten wertlos, weil Identitätsklau nicht verhindert wird.

In Photoshop ist die Handhabung der Technologie sehr intuitiv. Im Fenster „Content Credentials (Beta)“ kann man einstellen (Abb. 3.9) und einsehen (Abb. 3.10), welche Informationen gespeichert werden. Es lassen sich auch Social-Media-Konten verknüpfen. Beim Export hat man die Wahl zwischen dem Einfügen der CCr direkt in die Bilddatei oder/und dem Hochladen in die Content Credentials Cloud[11]. PS-Nutzer:innen müssen keinen Schlüssel erstellen, haben keinen Zugriff auf ihr Zertifikat, dementsprechend ist die Kontrolle dessen an das Adobe Cloud-Konto geknüpft. Als Inhaber:in einer Education-Lizenz steht einem die Funktion derzeit nicht zur Verfügung.

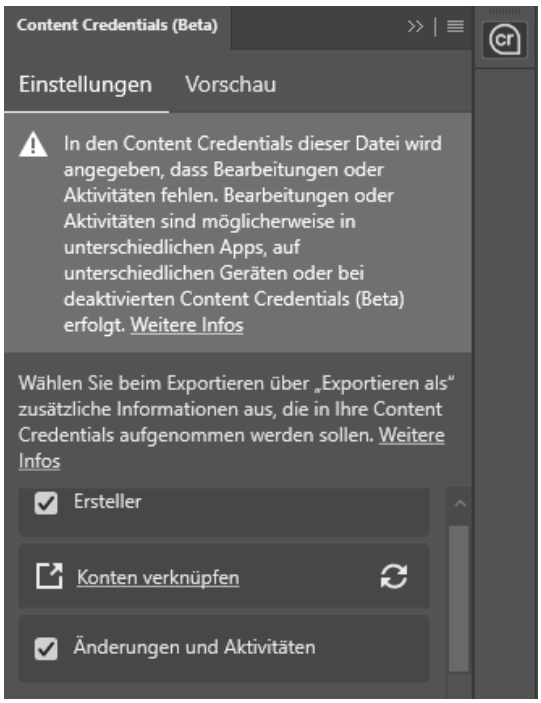


Abbildung 3.9: Ansicht in Photoshop nach Aktivierung der Funktion „Content Credentials (Beta)“

Die Erkenntnisse der Versuche fließen auch in Kapitel 5 und Kapitel 6 ein. Grundsätzlich ist festzuhalten, dass die Manifeste mehr Informationen enthalten, als auf der Verify-Seite zugänglich gemacht werden.

Es sei an dieser Stelle transparent gemacht, dass für die Versuche Adobe-Implementationen genutzt wurden, weil diese am fortgeschrittensten sind und andere Implementationen darauf aufbauen werden, weil Adobe das *Software Development Kit* und das c2patool öffentlich zur Verfügung gestellt hat. Entsprechend wurde auch ausschließlich die Verify-Seite der CAI verwendet, um eine möglichst schlüssige Darstellung und einen guten Vergleich erzielen zu können. Weil die Verify-Seite nur bedingt fundierte Einschätzungen hinsichtlich der Authentizität eines Bildes zulässt, lohnt sich die folgende Betrachtung bildforensischer Methoden in Kapitel 4.

[11] Die Content Credentials Cloud wird als öffentlicher und dauerhafter Speicher für CCr beworben (vgl. Inc. 2025).

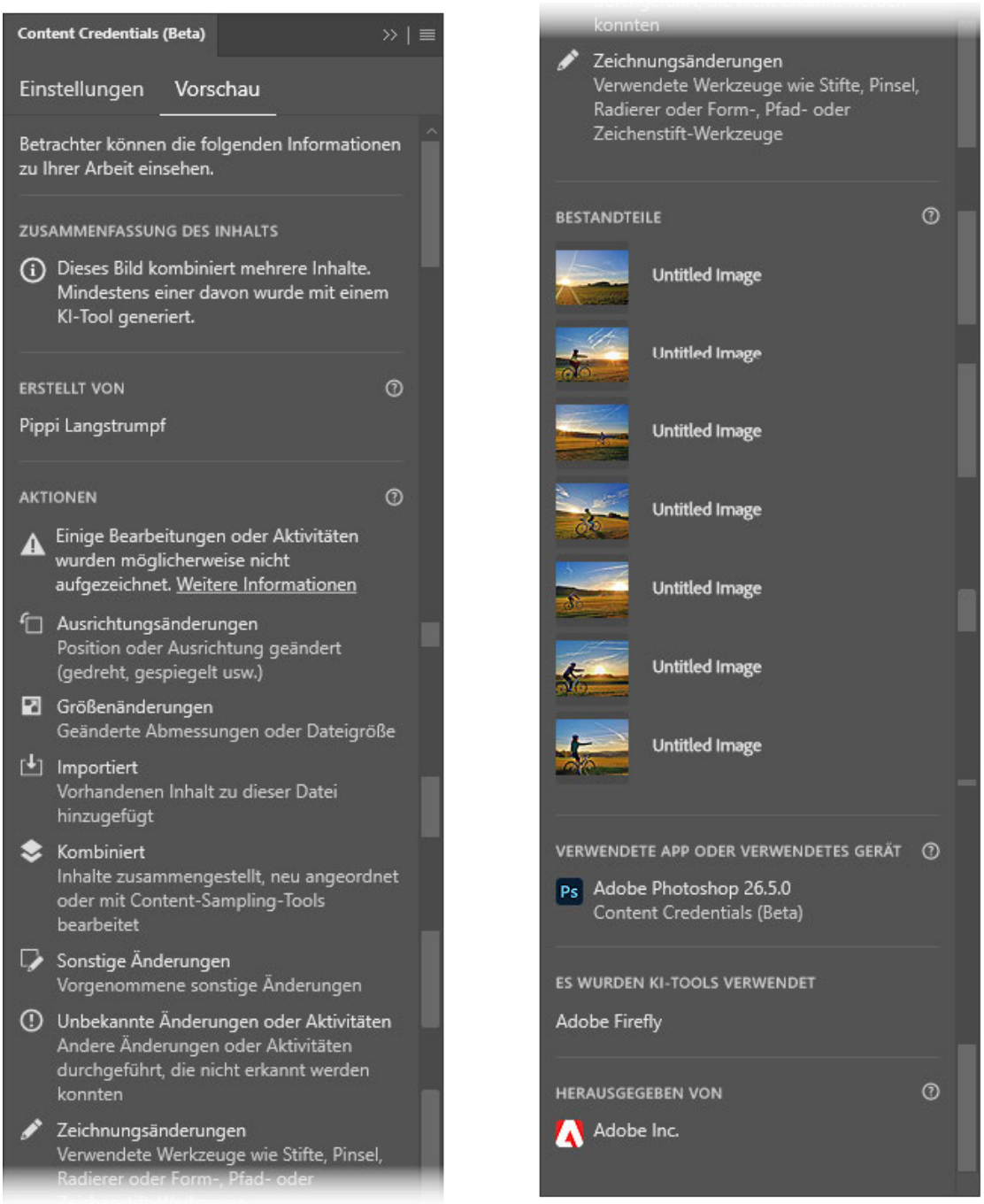


Abbildung 3.10: Vorschaufenster „Content Credentials (Beta)“ in Photoshop (aus Versuch 4, `V4-edit3.jpg`)

4 Bildforensik

Als Teil der Multimediaforensik (vgl. Gerling 2022, S. 305) kann Bildforensik auch als Teilgebiet der Informatik angesehen werden. Während Forschung in diesem Bereich überwiegend durch Informatiker:innen bzw. Computer Scientists betrieben wird[1], erfolgt die Anwendung der entwickelten Methoden in den Bereichen Kriminalistik und Journalismus, sowie in interdisziplinären Forschungsfeldern.
Die bekanntesten Methoden der Bildforensik stützen sich auf die Bildebene - also das für Menschenaugen Sichtbare im Bild (Abschnitt 4.1). Von bildforensischen Methoden auf Pixelebene (Abschnitt 4.2) spricht man, wenn nicht direkt sichtbare Bereiche durch Bearbeitung sichtbar gemacht werden, um Aufschluss über mögliche Bearbeitungen und Unstimmigkeiten zu erhalten. Folgt man den Ausführungen von Piva u. a. (2022), Krawetz und Xiang u. a. (2021), gibt es darüber hinaus noch Methoden auf Containerebene (Abschnitt 4.3). Zu den *unsichtbaren* Informationen einer Bilddatei zählen neben Metadaten auch Metainformationen, die aus der Struktur und dem Aufbau der Datei gewonnen werden können.
Bildforensische Untersuchungen haben zum Ziel, die Integrität und Authentizität eines Bildes zu beurteilen, und Manipulationen zu erkennen. Dabei wird Authentizität strenger und auf anderer Ebene definiert, als es die C2PA in ihrer Spezifikation vornimmt. Für einen direkten Vergleich sei erneut auf den Abschnitt 5.1 ab Seite 46 verwiesen.
Um Manipulationen zu erkennen, ist es „extrem hilfreich zu wissen, wie solche Fakes eigentlich hergestellt werden“ (Welchering 2020, S. 27), sich also mit aktuellen Möglichkeiten der Bildbearbeitung auszukennen (vgl. SWGDE 2025). Während in der journalistischen Praxis durch Bilduntersuchungen eine Beurteilung getroffen werden soll, betont die Scientific Working Group on Digital Evidence (SWGDE), dass durch

[1] Auf Grund der schnellen Entwicklung von KI beschäftigen sich zunehmend auch KI-Forscher:innen mit Möglichkeiten, diese für forensische Analysen einzusetzen.

die Analyse lediglich eine *Meinung* entstehen kann, also keine absolute Gewissheit bzw. Aussage erzeugt wird (vgl. Abb. 2 in SWGDE 2025). Für den Prozess ist die Unvoreingenommenheit der Forensiker:innen von großer Bedeutung, um den Bestätigungsfehler zu vermeiden, und möglichst objektiv analysieren zu können.

4.1 Bildebene

Riess beschreibt physik-basierte Methoden als Vorgehensweisen, die untersuchen, ob der Bildinhalt den Regeln der Physik folgt und somit als authentisch interpretiert werden kann (übersetzt nach Riess 2022, S. 209). Er unterscheidet nochmals zwischen geometrischen und photometrischen Methoden, sowie die Kombination beider.
Geometrische Methoden untersuchen, wie 3D-Objekte nach geometrischen Regeln auf einem 2D-Bild abgebildet werden müssten. Bei Unterschieden in der Projektion zweier Objekte im selben Bild wird von einer Bearbeitung ausgegangen. So können etwa nachträglich hinzugefügte Personen erkannt werden, sofern ihre tatsächliche Größe bekannt ist, indem die Größe der Abbildung in Relation zu den Abbildungen der anderen Personen oder bekannter Objekte (etwa Gebäude, Autos, Infrastruktur) gesetzt wird (vgl. Riess 2022, S. 210ff.). Riess beschreibt mathematisch, was vereinfacht ausgedrückt heißt: Die Proportionen müssen stimmen. Auch Verzerrungen durch seitliche Aufnahmen müssen konsistent sein und können durch Vergleichsobjekte im Bild überprüft werden (Homographie). Während der manipulierte Text auf einem seitlich aufgenommenen Papp-Schild auf den ersten Blick vielleicht noch „stimmig“ erscheint, kann ein geometrischer Vergleich zwischen der Verzerrung des Schilds und der Verzerrung der Schrift die Manipulation aufdecken[2]. Geometrische Untersuchungen setzten entsprechend voraus, dass ein Teil des Bildes nicht bearbeitet ist und als Referenz für geometrische Berechnungen dienen kann. Außerdem wird von bekannten Projektionsregeln, wie sie durch Linsen in Kameras gegeben sind, ausgegangen.
Bei einer Außenaufnahme mit blauem Himmel kann davon ausgegangen werden, dass die Sonne die einzige, zumindest aber die dominante Lichtquelle ist, folglich alle Schatten in die gleiche Richtung zeigen sollten und dessen Länge mit der Position und Größe der ver-

[2] Dieses sehr spezielle Beispiel stammt aus (Riess 2022, S. 219). Zeitungsberichte wie der von Russezki u. a. (2019) mit ähnlichen Szenarien unterstreichen jedoch die Relevanz.

ursachenden Objekte korrelieren muss. Ist das (vermeintliche) Aufnahmedatum bekannt, kann mit Hilfe von OSINT Werkzeugen wie beispielsweise `www.sonnenverlauf.de` die Richtung und Länge des Schattens sowie der Sonneneinfallswinkel approximiert werden. Durch Kamerablitz, komplexere Beleuchtungsszenarien und starke Nachbearbeitungen kann es jedoch zu Kontroversen kommen wie im Falle des World Press Photo of the Year 2013 (siehe hierzu Steadman 2013). Zwar ist die mathematisch gestützte Analyse[3] gut erforscht, jedoch ist eine manuelle Analyse zeitaufwendig und daher im journalistischen Kontext nicht zu erwarten.
Bei rein physik-basierten Methoden werden keine Informationen über den Veröffentlichungskontext berücksichtigt, der Wahrheitsgehalt einer Bildaussage kann damit allein folglich nicht bestimmt werden. Jedoch können physik-basierte Methoden auch bei schlechter Auflösung oder geringer visueller Qualität angewandt werden, was im Vergleich zu folgenden Methoden als Vorteil gewertet werden kann (vgl. Riess 2022, S. 207). Der Aufnahmeort und Zeitpunkt kann durch Analysen auf Bildebene approximiert oder sicher ermittelt werden. Zur Authentifizierung werden die einzelnen Bildelemente analysiert und geprüft, ob sie zusammenpassen. So werden etwa sichtbare Schrift, Autokennzeichen, Straßenschilder, aber auch Flora, Fauna sowie Kleidung und das Aussehen vorhandener Menschen im Bild untersucht. Dadurch lässt sich etwa schlussfolgern, in welchem Land und zu welcher Jahreszeit das Bild aufgenommen wurde. Derartige Analysen sind zeitaufwendig und erfordern Hintergrundwissen, können jedoch bei fehlenden Metadaten die Authentifizierung und Verifikation ermöglichen.

Noch werden KI-generierte Bilder durch einen genauen Blick auf direkt sichtbare Details identifizert. Gröbere Details wie Hände, Proportionen von Objekten zueinander und ineinander verschlungene, oder überlappende Elemente werden noch schlecht durch KI dargestellt, aber Experten verweisen bereits auf feinere Details wie Haut- und Haarstrukturen und „skin-to-object contact“ (SWGDE 2025, S. 4). Die Details können durch aufwendige Nachbearbeitung oder durch das Voranschreiten der KI-Entwicklung verbessert werden, sodass mittelfristig davon ausgegangen werden muss, dass eine visuelle Analyse allein nicht mehr zur Erkennung von synthetischen Bildern ausreicht. Noch sind KI-generierte Bilder wesentlich „sauberer“ als Fotografien. Staubpartikel in der

[3] Mit der mathematisch gestützten Analyse beschäftigen sich laut Riess viele Publikationen, namentlich er selbst und die Autoren Farid, Kee und Peng.

Luft oder optische Verzerrungen müssen mittels Prompt explizit verlangt werden. Nichtbearbeitete Fotografien enthalten konsistente Hinweise auf die Verwendung einer Kamera; auf Bildebene kann die Tiefenschärfe ermittelt und mit Metadaten abgeglichen werden (vgl. SWGDE 2025, S. 8); auch das Bokeh einer Kamera ist spezifisch und lässt Rückschlüsse auf die verwendete Linse zu. Bewegungsunschärfe deutet auf eine Fotografie hin, dadurch lassen sich auch Bewegungsvektoren ermitteln und auf Plausibiliät prüfen. Ziel der Untersuchung der direkt sichtbaren Szene eines Bildes ist es, eine mögliche Bearbeitung zu erkennen (vgl. SWGDE 2025, S. 8). Weil generierte Gesichter bereits nicht mehr von echten Gesichtern zu unterscheiden sind, ist allein durch Analyse des Bildinhaltes keine Aussage, auch keine Meinung darüber möglich, ob ein Bild synthetischen Ursprungs ist, oder nicht. Es müssen ergänzend Analysen auf Pixel-und Containerebene erfolgen.

4.2 Pixelebene

Wenn mann nicht die Elemente im Bild, sondern Pixelstrukturen wie Kompressionsartefakte, Bildrauschen und Helligkeitsverteilung analysiert, spricht man von forensischen Methoden auf Pixelebene (vgl. Xiang u. a. 2021). Dabei werden statistische Unregelmäßigkeiten und Artekfakte durch digitale Bearbeitung des Bildes sichtbar gemacht. Für die richtige Interpretation bedarf es auch Wissen aus der Informations- und Medientechnik. Jeder Kamerachip ist durch Rauschmuster, Farbwiedergabe, Sensorreaktion und Pixelfehler einzigartig. Es existieren Verfahren, mit denen sich der dadurch entstehende „Fingerabdruck“, auch als Photo Response Non-Uniformity Pattern (PRNU) bezeichnet, einer einzelnen Kamera auslesen lässt. Vergleicht man diesen mit dem Fingerabdruck eines fraglichen Bildes, kann überprüft werden, ob das Bild mit eben dieser Kamera aufgenommen wurde (vgl. Gerling u. a. 2018, S. 166; Lorch 2023; Kirchner 2022). Dazu bedarf es allerdings der Originaldatei und der infrage kommenden Kamera. Schon einfache Bearbeitungsschritte und Kompressionen führen zum (teilweisen) Verlust des Fingerabdrucks (vgl. Butora u. a. 2024). Die stetig wachsende Bildverarbeitungskette

von Smartphone-Kameras[4] erschwert die Analyse von PRNU, weil automatisierte Verbesserungen wie Verzerrungskorrektur, Bildstabilisation und die Kombination mehrere Aufnahmen zu einer Bilddatei den Fingerabdruck unkenntlich machen können (vgl. Kirchner 2022, S. 66ff). Dennoch findet diese Analysemethode gerade in kriminalistischen Fällen Anwendung (vgl. SWGDE 2025).

„Kameras (...) komprimieren anders als Bildbearbeitungsprogramme“ (Gerling 2022, S. 309). Bei genügend großen Datensätzen lassen sich spezifische Kompressionsmuster und Quantisierungstabellen[5] feststellen und vergleichen. Sind verschiedene Kompressionsartefakte in einem Bild vorhanden, deutet das auf eine Komposition, mindestens aber auf mehrfache Bearbeitung hin (vgl. Lorch 2023, S. 60). Die *Error Level Analysis* (ELA)-Methode nutzt genau diesen Umstand, um Unregelmäßigkeiten sichtbar zu machen; sie kann auf `https://29a.ch/photo-forensics` getestet werden. Auch hier wird die Analyse durch starke Nachbearbeitung erschwert, insbesondere wenn KI zum Einsatz kam.

Weitere Methoden sind *Noise Inconsistency Detection*, bei der unterschiedliche Rauschsignaturen innerhalb eines Bildes auf eine Komposition hinweisen können, sowie die Analyse von Interpolationsartefakten. Das Chrome Plug-In „Fake news debunker by InVid & WeVerify“ vereint viele Methoden zur Bild- und Videoanalyse. Es wurde im Rahmen verschiedener europäischer Projekte seit 2016 stetig weiterentwickelt und steht öffentlich zur Verfügung. Es stehen dort verschiedene Analysewerkzeuge zur Verfügung: GHOST vergleicht Helligkeitsverteilungen im Bild mittels Histogramm-Analyse; DCT analysiert die DCT-Koeffizienten, deren Anomalien auf hinzugefügte Assets hindeuten können und die WAVELET-Analyse ermöglicht die genauere Untersuchung feiner Strukturen.

Die einzelnen Bildanalyse-Verfahren können Indizien für mögliche Bildmanipulation sichtbar machen, deren Ergebnisse müssen jedoch stets im jeweiligen Kontext und mit tiefgehendem technischen Verständnis interpretiert bzw. bewertet werden. Das macht den Prozess stark erfahrungsabhängig und zeitaufwendig.

[4] Gemeint ist Postprocessing in der Kamerapp, wodurch ohne Einflussnahme der Nutzer:innen *vor* dem Speichervorgang bereits weitreichende Bearbeitungen erfolgen. Schon 2016 wurde bei der Vorstellung des Apple iPhones 7 beschrieben, dass der *Image Signal Processor* pro gespeichertes Bild über 100 Milliarden Prozesse durchführt (vgl. *Apple - September Event* 2016).

[5] Vgl. hierzu Abschnitt 4.3.

4.3 Containerebene

Selbst wenn auf Bildebene und Pixelebene keine Anzeichen für Manipulation vorliegen, so können die Dateistruktur und insbesondere die in ihr gespeicherten Metadaten mehr über den Entstehungsprozess des Bildes verraten.

> „However, any methodology applied to image authentication should incorporate both image content and image structure.“ (SWGDE 2025, S. 5)

Piva u. a. (2022) sprechen von *file container analysis*; es wird ausgenutzt, dass der Aufbau einer Bilddatei nicht vollständig standardisiert ist und bestimmte Kombinationen von APP-Segmenten, IFDs und deren Reihenfolge auf bestimmte Hersteller oder Software hindeuten können.

In den vergangenen 25 Jahren wurden Stück für Stück mehr Merkmale (*features*) der Bilddatei genutzt, um eine Art Bildsignatur[6] zu erstellen, die dann mit großen Datensätzen automatisiert, und in jüngster Zeit auch KI-gestützt, verglichen werden kann. Piva und Iuliani beschreiben mehrere Ansätze und vergleichen sie in ihrer Effektivität und Zuverlässigkeit. So konnte Kee Eric bereits 2011 mit Hilfe von 284 *header features* das Bild mit einer Sicherheit von 69,1 % genau einem Kameramodell zuordnen (vgl. Piva u. a. 2022, S. 369ff). Nutzt man neben den Exif-Metadaten noch die Quantisierungstabelle(n) eines Bildes, kann bei iPhone-Fotos mit fast 82%-iger Wahrscheinlichkeit richtig auf die iOS-Version geschlossen werden (vgl. Piva u. a. 2022, S. 371ff). Auch verwendete Software und Social-Media-Plattformen hinterlassen eindeutige Spuren, was auch mit Hilfe der eigenen Versuche belegt werden konnte (vgl. Abschnitt 3.2).

> „[W]e are currently not aware of any publicly available software that would allow users to consistently forge such information without advanced programming skills. This (...) reemphasizes that file characteristics and metadata must not be dismissed as unreliable source of evidence for the purpose of file authentication per se.“ (Piva u. a. 2022, S. 364)

Die Forscher:innen kommen zur klaren Konklusion, dass Metadaten nicht per se igno-

[6] Das selbe Vorgehen soll auch zur Wiederherstellung von CCr benutzt werden, wird dann als *fingerprinting* bezeichnet, siehe hierzu Seite 16.

riert werden dürfen. Das alles setzt jedoch voraus, dass die *features*, insbesondere die Metadaten nicht gelöscht wurden.

Sind in einer Bilddatei Exif-Daten vorhanden, eventuell sogar kameraspezifische Metadaten wie in Versuch 1 im Bild `V1.jpg`, so ist das ein Indiz für eine digitale Fotografie. Auch der Dateiname kann bereits viel verraten, vorausgesetzt er wurde nicht geändert; Kamera- und Kameraapp-Hersteller erzeugen sie nach bestimmten Mustern, oft ist auch das Aufnahmedatum Teil des Dateinamens. KI-generierte Bilder enthalten keine Exif-Metadaten (vgl. Abschnitt 3.3). Ob ein Bild eine unbearbeitete digitale Fotografie ist, also gewissermaßen direkt aus der Kamera stammt (*directly photographed*) lässt sich jedoch nur über das Ausschlussverfahren mit Sicherheit erörtern. „Forensisch authentifiziert ist ein solchermaßen examinierter Datensatz, wenn es nicht gelingt, seinen Strukturmerkmalen informationstechnische Inkonsistenzen nachzuweisen“ (Rothöhler 2024, S. 39). Auch Gerling beschreibt, dass Bilder direkt aus der Kamera authentischer wahrgenommen werden, als sichtlich bearbeitete Bilder (vgl. Gerling 2022, S. 3). Somit ist es für die Beurteilung der Authentizität eines Bildes wichtig, durch Metadatenanalyse eine Bearbeitung eindeutig ausschließen oder feststellen zu können.
Finden sich in den Metadaten Hinweise auf verwendete Software, wie etwa in den Einträgen `XMP-xmp:CreatorTool` oder `IFD0:Software`, oder gar Software-spezifische Kategorien wie `XMP-GIMP`, `Photoshop`, `XMP-microsoft` oder `Adobe`, so muss davon ausgegangen werden, dass das betrachtete Bild bearbeitet wurde. Auch zusätzliche Marker wie `APP0` oder das Fehlen von Miniaturbild-Informationen in den `IFD1`-Einträgen ist ein Hinweis auf eine bearbeitete Fotografie (vgl. Gangwar u. a. 2018, S. 340). Letzteres lässt sich auch in Versuch 4 in den Metadaten von `V4-edit1.jpg` im Vergleich zum Original `V4.jpg` beobachten. Werden hingegen nur einzelne Metadaten-Einträge bearbeitet, so bleiben Miniaturbild-Informationen erhalten (vgl. Abschnitt 3.1). Sind keine oder nur sehr wenige Metadaten vorhanden, kann das auf eine Bildbearbeitung hindeuten, wahrscheinlicher ist jedoch, dass diese absichtlich gelöscht, oder bei Veröffentlichung im Internet automatisiert gelöscht bzw. verändert wurden. Sind Einträge in der Kategorie `XMP-xmpMM` [7] oder `CBOR:ActionsAction` [8] vorhanden, kann direkt auf die Art der Bearbeitung geschlossen werden. Sind `IPTC`-Einträge vorhan-

[7] Sehr deutlich in `V1-edit3.jpg` zu sehen, wo GIMP vermerkt, dass lediglich Metadaten geändert wurden.

[8] Vergleiche hierzu `V4-edit3.jpg` und dort das Feld `CBOR:ActionsAction` mit 18 Einträgen.

den, wurden Metadaten bewusst durch einen Menschen bearbeitet; Rückschlüsse auf Veröffentlichungskontext und Fähigkeiten der Person sind somit möglich.

Auch Plattformen, auf denen Bilder hochgeladen werden können, hinterlassen Spuren in den Metadaten, wie Versuch 5 belegt hat. Facebook und Instagram erzeugen einen Eintrag `IPTC:SpecialInstructions` und fügen als Wert eine Ziffernfolge beginnend mit `FBMD` [9] hinzu. Auch ein kryptisch wirkender, langer Dateiname deutet darauf hin, dass das Bild von einer Plattform heruntergeladen, oder über solche geteilt wurden. Via WhatsApp geteilte Bilder erhalten einen Dateinamen mit Präfix `WA`, via Signal geteilte Bilder das Präfix `signal`. Allgemein löschen Online Plattformen die meisten Metadaten, um Speicherplatz zu sparen und die Privatsphäre der Nutzer:innen zu schützen. Das führt zum Verlust möglicher Herkunftsnachweise. „Ohne EXIF- oder ähnliche Metadaten ist das Wissen um den Ursprung eines Bildes gering“ (Gerling 2022, S. 9).

Die Möglichkeiten, Ursprung und Zustand von Bildern durch forensische Mittel herauszufinden, sind vielfältig. Dabei haben Untersuchungen auf unterschiedlichen Ebenen jeweils ihre Vor- und Nachteile. Eine fundierte Analyse beinhaltet deshalb stets Methoden auf allen drei Ebenen und gewichtet diese der Beschaffenheit des Bildmaterials entsprechend. Ziel ist eine möglichst objektive Einschätzung, jedoch keinesfalls eine eindeutige Beurteilung im Hinblick auf die Authentizität des geprüften Materials. Metadaten spielen auf Bild- und Pixelebene eine komplementäre, auf Containerebene jedoch eine entscheidende Rolle.

[9] Vermutlich steht das für „Facebook Metadata“.

5 Einfluss der C2PA-Spezifikation

Die vorangestellten Kapitel warfen einen theorie-fokussierten Blick auf Elemente und Methoden, die bei der Authentifizierung hilfreich sein können. In diesem Kapitel werden die potenziellen Auswirkungen der C2PA-Spezifikation erörtert. Es bedarf zunächst einer genauen Beschreibung des Konstruktes Authentizität. Danach folgt ein Blick auf die Chancen, Risiken und Schwachstellen der Spezifikation, um in Kapitel 6 eine Synthese aller betrachteten Aspekte vornehmen zu können.

5.1 Authentizität

Der Begriff wird kontextbezogen unterschiedlich verwendet und interpretiert. Ein besonderes Kriterium bei der Unterscheidung zwischen authentisch und nicht-authentisch ist die nachweisbare oder auch nur wahrgenommene Diskrepanz zwischen Original und bearbeiteter Version.

5.1.1 Verschiedene Blickwinkel auf den Begriff

Das Wort *authentication* lässt sich übersetzen mit Authentifizierung, Beglaubigung, Anerkennung, Echtheitserklärung, Bestätigung; ist etwas oder jemand authentifiziert, so ist es oder er:sie authentisch, echt, man kann es oder ihm:ihr glauben. Das Wort *provenance*[1] wird durch C2PA oft zusammen mit *authenticity* verwendet, ist aber nicht das Selbe.
Was im allgemeinen Sprachgebrauch unter Authentizität verstanden wird, lässt sich gut

[1] Zur Begriffserörterung sei auf Abschnitt 2.2.1 auf Seite 10 verwiesen.

in sozialen Netzwerken beobachten. Salisbury u. a. (2017) identifizieren verschiedene Marker, die Nutzer:innen dazu bewegt, einen medialen Inhalt oder ein Profil als authentisch wahrzunehmen. Dazu zählen unter anderem: Konsistenz (zwischen der Online-Darstellung und dem offline erlebten Menschen), Spontaneität (Gestelltes wirkt nicht authentisch), und Amateurismus (je weniger Bearbeitung ersichtlich ist, desto authentischer wirkt das Bild (vgl. Salisbury u. a. 2017). Gerling fügt hinzu: „Was schnell geht, kann nicht gefälscht sein“ (Gerling 2024, S. 2), entsprechend kann auch eine geringe Zeitspanne zwischen Erstellung und Veröffentlichung eines Bildes ein Argument für die Authentizität eines Bildes sein (ebd.). Es ist verständlich, dass Forensiker:innen und Journalist:innen andere Auffassungen zum Begriff haben, da die Content Credentials aber als Authentifizierungs-Werkzeug für jede und jeden beworben werden, ist auch die Interpretation und Verwendung des Begriffs durch die gesamte Bevölkerung relevant.
Für Forensiker:innen ist ein Bild authentisch, wenn keine Bearbeitungen erkennbar sind, welche die Regeln der Physik verletzen (vgl. Kapitel 4). Piva und Iuliani gehen noch einen Schritt weiter: „the analysis of EXIF data structures can be useful to discriminate between authentic and edited images“ (Piva u. a. 2022, S. 371). Das liest sich so, als würden für die Autor:innen nur gänzlich[2] unbearbeitete Bilder als authentisch gelten. Die Bildforensik versucht möglichst nachvollziehbar und damit objektiv zu einer Einschätzung hinsichtlich der Authentizität eines Bildes zu gelangen. Im Gegensatz dazu zählt auf Social-Media-Plattformen das individuelle Urteil.
Auch der Philosophie-Professor Gerth beschreibt „Bildauthentizität als subjektive Zuschreibung von Glaubwürdigkeit zu einer Fotografie“ und unterscheidet klar zwischen dieser und der „Visuellen Wahrheit“ (Gerth 2018, S. 7). Demnach ist Authentizität „ein situiertes und veränderbares soziales Konstrukt“ (Krämer u. a. 2018, S. 17) das je nach Kontext und Publikum unterschiedlich[3] interpretiert werden kann. Wenngleich genannte Autor:innen beschreiben, dass die Verwendung des BegriffsAuthentizität stark kontextabhängig ist, so sind sie sich doch einig, dass im Bezug auf visuelle Authentizität die Realitätsnähe des Gezeigten entscheidend ist. Gerade im journalistischen Kontext sollen Bilder Wirklichkeit vermitteln; Rezipient:innen nutzen Bilder, um auch ohne eigene Erfahrungen ein umfassendes Weltbild zu konstruieren (Gerth 2018, S. 13) und Haltungen

[2] Gänzlich meint hier sowohl auf Bild-, Pixel-, als auch auf Containerebene (vgl. Kapitel 4).
[3] Die Kommunikationswissenschaftler:innen merken zudem an, dass dieses Konstrukt auch kulturell beeinflusst sein kein, es dazu aber noch keine brauchbaren Untersuchungen gibt (vgl. Krämer u. a. 2018, S. 17).

zu entwickeln.
Weil nahezu alle Pressefotos heutzutage vor Veröffentlichung bearbeitet werden (vgl. Kübler 2022, S. 319), gäbe es aus bildforensischer Perspektive nur sehr wenige authentische journalistische Bilder. Grittmann definiert den Begriff als eine „auf sozialen Praktiken und professionellen Normen beruhende Konstruktion von Wirklichkeit“ (zitiert nach Gerling 2022, S. 3, ursprünglich Grittmann 2003, S. 125) und zieht damit Parallelen zur geisteswissenschaftlichen Auffassung des Begriffs. Zwar schreiben (Foto)-Journalist:innen selten über ihre eigene Authentizität oder die ihrer Bilder, vor dem Hintergrund der journalistischen Aufgabe, die Öffentlichkeit wahrhaftig zu unterrichten (vgl. Deutscher Presserat 2025) ist deren Auffassung des Begriffs jedoch ebenfalls relevant. Bildredaktionen entscheiden, welche Bildbearbeitung unternehmensintern zulässig ist, und welche nicht.

5.1.2 Die CAI erweitert das Konstrukt

Simmons und Winograd interpretieren *provenance* und *authenticity* auf Grundlage der C2PA-Spezifikation stets zusammen und knüpfen Authentizität an den Willen der Quelle: „authenticity refers to whether the content has been manipulated or altered in a way out of the control of the trusted source of the information“ (Simmons u. a. 2024, S. 1). Nach dieser Auffassung ist ein Bild authentisch, wenn es genau so ist, wie es der:die (vertrauenswürdige) Urheber:in veröffentlicht hat.
Die NSA u.a. (2025) formulieren und interpretieren vorsichtiger; Informationen über Herkunft und Entstehungsprozess können gut informierte Entscheidungen[4] erleichtern, jedoch keine Aussage über die Glaubwürdigkeit des Bildinhaltes liefern. Nichtmanipulierte Metadaten können zur Authentifizierung von Bildern genutzt werden (vgl. Gerling 2022, S. 299) und die Anwendung der C2PA-Spezifikation kann dies unterstützen, vorausgesetzt die Metadaten sind nicht manipuliert.
Auch KI-generierte Bilder können nach der technischen Interpretation der C2PA authentifiziert werden, und damit als authentisch wahrgenommen werden. Der Information Technology Industry Council (ITI) beschreibt Authentifizierung im Kontext von KI als „act of verifying or confirming the authenticity of content generated by AI models“ (ITI

[4] Im Original: „informed descisions“ (NSA u.a. 2025, S. 4ff).

2024, S. 18). C2PA-Assoziierte schreiben offensiv, dass sie mit CCr, Wasserzeichen und *fingerprinting* versuchen, robustes Vertrauen[5] zu ermöglichen, „to support stories told by real and generative images“ (Collomosse u. a. 2024a, S. 1).

KI-generierte Bilder als „authentisch“ bzw. „authentifizierbar“[6] darzustellen, steht im starken Kontrast zur geisteswisschenschaftlichen Auffassung. Wenn Bilder Wirklichkeit vermitteln sollen, dann können synthetische Bilder nicht „authentisch“ genannt werden, folglich allenfalls technisch authentifiziert werden. Ein Graubereich sind synthetische Bilder, die eine reale Situation nachbilden, etwa mit dem Ziel Persönlichkeitsrechte zu schützen. Hier kann die Bildaussage authentisch sein, jedoch kann nicht von einer authentischen Fotografie, oder allgemeiner von einem authentischen Bild gesprochen werden. Die Betrachtung deutet auf eine Verschiebung hin: Der Fokus im Authentifizierungsprozess rückt mit C2PA vom Bildinhalt ab, hin zur technischen Ebene mit Blick auf den Entstehungsprozess. Sobald dieser anhand von *provenance*-Informationen nachvollzogen werden kann, spricht die C2PA von authentifizierbaren Bildern. Während Forensiker:innen, Journalist:innen, Geisteswissenschaftler:innen und auch die Allgemeinheit Authentizität größtenteils auf Bildebene wahrnehmen, prüft die C2PA nur auf technischer Ebene. Dass CCr lediglich einen Indikator liefern sollen, um die Authentifizierung zu erleichtern (vgl. C2PA V2.1, Abschnitt 1.1), jedoch keine Aussage im Bezug auf den Inhalt eines Bildes treffen, bleibt zu oft unerwähnt oder wird nicht genügend betont.

[5] Im Original: „robust trust signals“ (Collomosse u. a. 2024a, S. 1).

[6] Der Neologismus ist angelehnt an den Begriff *identifizierbar*. Etwas lässt sich nur authentifizieren, wenn es eine Authentizität besitzen kann. Dann kann es authentifiziert werden, es ist authentifizierbar.

5.2 Chancen

Um durch KI erzeugte Inhalte[7] zu kennzeichnen gibt es auch andere Methoden, jedoch wäre die Implementierung von CCr sehr effizient und könnte automatisiert von Systemen erkannt werden. Zudem können dezidierte Informationen über das verwendete Model und den zum Training genutzten Datensatz[8] protokolliert werden. Der Eintrag `digitalSourceType` muss dem NewsCodes Scheme[9] der IPTC entsprechend gefüllt werden und gibt Auskunft, ob ein Bild eine direkte digitale Fotografie (`digitalCapture`), ein nicht-mit-KI-bearbeitetes Asset (`humanEdits`), mit KI bearbeitet (`compositeWithTrainedAlgorithmicMedia`), oder gänzlich KI-generiert (`trainedAlgorithmicMedia`) ist (vgl. *NewsCodes Scheme* 2024). In Kombination mit dem Feld `regionOfInterest` , welches den Bildausschnitt definiert, auf den sich eine Assertion bezieht, ließe sich beurteilen, ob die Aussage des Bildinhaltes durch die Bearbeitung verändert wurde[10]. Die Spezifikation ermöglicht diese Differenzierung prinzipiell, eine Anwendung konnte in den Versuchen (etwa bei `V4-edit3.jpg`) nicht festgestellt werden.
Somit könnte die systemweite Umsetzung der C2PA-Spezifikation die Erkennung KI-generierter und mittels KI veränderter Bilderr enorm erleichtert, und potenziellen Fake News entgegen gewirkt werden.

Verknüpft man Social-Media-Konten mit seinem Adobe-Creative-Cloud-Konto, so können Verweise auf jene in die CCr eingefügt werden, sodass Rezipient:innen über die Verify-Seite zu den Profilen der Künstler:innen weitergeleitet werden. So würde auch bei mehrfach geteilten Inhalten die Urheberschaft jederzeit einsehbar und auf den:die Künstler:in zurückführbar sein. Auch der Hinweis, dass ein Bild nicht für KI-Training genutzt werden soll, kann durch Kreative in den CCr vermerkt werden (vgl. NSA u.a. 2025, S. 10).

Es ist unbestritten, dass Citizen Media, also „Amateurvideos und Bilder“ (Ratering

[7] Im Englischen auch als with Artifical Intelligence Generated Content (AIGC) bezeichnet.
[8] Siehe hierzu Tabelle 11 in Abschnitt 18.16 der Spezifikation V. 2.1 auf Seite 184.
[9] `http://cv.iptc.org/newscodes/digitalsourcetype/`
[10] Wäre klar, dass mit Hilfe von KI beispielsweise lediglich ein Laternenpfahl aus dem Bild entfernt wurde, so müsste man die Gesamtaussage des Bildes nicht in Frage stellen, obwohl KI bei der Bearbeitung zum Einsatz kam.

2022, S. 1) immer mehr Präsenz in den Medien erlangen (vgl. zudem Gerling u. a. 2018; Okeowo 2022). Nicht-journalistische Inhalte werden verallgemeinernd auch als UGC bezeichnet. Werden diese „Aufnahmen von Beteiligten, das heißt von ‚zufälligen Zeugen'" (Gerling u. a. 2018, S. 12) oder auch ganz gezielt von Menschenrechtsaktivist:innen erstellt, so spricht man im Englischen und rechtswissenschaftlichen Raum von *Citizen Media* oder auch *Citizen Journalism* (vgl. Okeowo 2022; Gerling 2022, S. 22).
Citizen Media vermag es, Ereignisse und Ungerechtigkeiten aus dem globalen Süden in das Bewusstsein der westlichen Bevölkerung zu rücken, Investigationen und Gerichtsverfahren zu initiieren (vgl. Okeowo 2022) und Beweismittel zu sein. Citizen Media verhilft Ereignissen zu Aufmerksamkeit, bei denen keine professionellen Journalist:innen anwesend sind. Beispielsweise wurde das sogenannte Hasi-Video vom 26. August 2018 von einer Amateurin aufgenommen. Es führte zu einer gesellschaftlichen und politischen Debatte (siehe etwa Wolters u. a. 2018). Noch größer war der globale Aufschrei nach der Veröffentlichung der Handy-Aufnahmen von der Verhaftung und Tötung von George Floyd 2020 durch einen Polizisten.
Sowohl für Menschenrechtsaktivist:innen als auch Journalist:innen insgesamt und insbesondere in Krisensituationen hätte der systemweite Einsatz von C2PA-Metadaten in Smartphonekameras den Vorteil, dass eben diese Citizen Media schneller und sicherer authentifiziert, und somit in der Berichterstattung verifiziert werden könnten. Weil die C2PA von Beginn an auch die Rezipient:innen in ihren Überlegungen mitgedacht hat, könnten auch diese von den zusätzlichen, und niedrigschwellig verfügbaren Informationen durch CCr profitieren. Durch den CCr-Pin würden mehr Menschen die Herkunft eines Bildes mit fragwürdigem Inhalt überprüfen können, ohne sich dafür technisches Wissen aneignen, oder spezielle Software installieren zu müssen. So könnte der verantwortungsvolle Umgang mit Online-Medien gefördert werden. Außerdem würde es mehr Transparenz zwischen Nachrichtenunternehmen und Bürger:innen erzeugen, was das Vertrauen in jene stärken könnte.

5.3 Risiken

Neue Technologien bedeuten auch neue Risiken. In Form von Technikfolgeabschätzungen und Risikomanagement wird bereits im Entwicklungsprozess der Spezifikation ein

Blick auf mögliche Schäden, Missbrauch und Auswirkungen durch Fehlanwendungen geworfen. Mögliche Lösungsstrategien zur Risikoreduktion veröffentlicht die C2PA unter dem Stichwort *C2PA Harms Modelling* (vgl. C2PA 2025). Die Risikoanalyse ist auch in Form von Tabellen veröffentlicht (etwa hier: C2PA 2022).

Eine viel diskutierte Eigenschaft der C2PA-Spezifikation ist die Möglichkeit, das Manifest unabhängig vom Asset zu speichern und lediglich eine URL in das Manifest zu schreiben. Diese URL zeigt dann auf das in einer externen Datenbank gespeicherte Manifest. Während die Spezifikation und auch die Risikoanalyse klar auf *Hard Bindings*[11] setzt (vgl. C2PA 2022), wird mit der Kombination[12] aus Wasserzeichen, Signaturen und Manifestspeicher doch stark auf diesen Prüfweg über *side-cars* gedrängt. Für die Einsicht in die Informationen der CCr ist es folglich notwendig, dass Rezipient:innen einen Link öffnen. Dadurch können die Webseitenbetreiber:innen der *repositories* nachvollziehen, wer wann welche Inhalte aufgerufen hat. Die Verfolgung von Nutzer:innen im Netz würde so erleichtert werden.

Aus soziologischer Perspektive muss angemerkt werden, dass durch CCr eine Subjektivierung erfolgt; zukünftig soll jede:r eigenverantwortlich die Herkunft von Bildern überprüfen und daraus die Glaubwürdigkeit des Gezeigten ableiten. Auch die Verknüpfung von Asset und Manifest soll durch Rezipient:innen selbst überprüft werden. Durch diese „human-in-the-loop"-Taktik (C2PA 2022, S. 8ff) soll die Schwäche der *Soft Bindings* ausgeglichen werden. War es bisher Aufgabe der Medienunternehmen und Politik, die Wahrheit herauszufinden und zu vermitteln, fällt es mit CCr ein Stück weit auf das Individuum zurück. Dem ist entgegen zu halten, dass CCr auch als zusätzliche Kontrollfunktion in der Beziehung zwischen Bürger:innen und Medienunternehmen verwendet werden können.

In manchen Situation ist das Löschen[13] einzelner Assertions erwünscht, etwa wegen Fehleingaben in den IPTC-Metadaten, oder um Informationen zu einer gelöschten Ebene zu entfernen. Das Redigieren erfolgt entweder durch tatsächliches Löschen der Assertion oder des Manifests, oder die Daten werden durch Nullen ersetzt. Abschnitt 6.8 der Spezifikation V2.1 legt fest, dass dabei zusätzliche Claims erstellt werden müssen, die das Redigieren nachvollziehbar machen. Durch die URI ist erkennbar, welche Art von Infor-

[11] Für eine Begriffserklärung siehe Abschnitt 2.2.2.
[12] Collomosse u. a. (2024a) nennen die Kombination *durable Content Credentials*, vgl. Seite 15.
[13] Der Prozess wird Redigierung (*redaction*) genannt und in Abschnitt 6.8 der Spezifikation erläutert.

mation gelöscht wurde. Es dürfen keine *Action-Assertions* und keine *Hash-Assertions* redigiert werden. Durch den Einsatz von Distributed Ledger Technology (DLT) würde das Redigieren von Assertions und damit von versehentlich in Umlauf gebrachten sensiblen Daten verunmöglicht (vgl. Rathi u. a. 2024), weil DLT kein löschen oder ändern zulässt (vgl. C2PA V2.1, Abschnitt 9.1). Die Risikoanalyse-Tabelle der C2PA spricht sich gegen[14] die Verwendung von DLT aus (vgl. C2PA 2022, S. 6). Statt dessen sollen Anwender:innen darauf hingewiesen werden, dass die Einbeziehung von Informationen und das Verwenden der CCr auf Freiwilligkeit beruht.

Das Offenlegen von Informationen zu Herkunft und Bearbeitungshistorie erfordert ein Abwägen zwischen der dadurch erzeugten Transparenz und eventuell notwendigem Schutz persönlicher Daten, Individuen und kreativer Prozesse. Es ist insbesondere dann schwierig, die Identität der Quelle zu schützen, wenn durch den Einsatz C2PA-fähiger Kameras[15] eine Seriennummer, Geodaten oder andere aufnahmespezifische Metadaten gespeichert werden, die auf ein Individuum und dessen Aufenthaltsort oder Bewegungsmuster hindeuten. Die Risikoanalyse führt unter „Infringement on human rights“ auch den Punkt „Privacy loss“ (C2PA 2022, S. 5) auf und sieht eine mögliche Entschärfung der Risiken dadurch gegeben, dass der Einsatz von C2PA keinerlei sensible Daten verlangt, die Art und Menge vermerkter Daten von Anwender:innen selbstständig ausgewählt werden kann und insgesamt auf Freiwilligkeit beruht (*opt-in*), und ein anonymes bzw. pseudonymes Signieren mittels „W3C credentials“ möglich ist (vgl. C2PA 2022, S. 5ff). Dem gegenüber steht jedoch das folgende Risiko.

Die epistemische Ungerechtigkeit, gekoppelt mit sich durch C2PA verfestigenden Machtstrukturen stellt wohl das größte Risiko der Technologie dar.

> „Another concern is one that was described under our assessment — epistemic injustice — which essentially addresses how [this could] strengthen existing relationships of power. We can imagine that this sort of system could be potentially implemented by social media for algorithmic ranking. So if you do include a C2PA manifest into an image on Twitter [jetzt **X**, Anm. d. Autorin], then it is possible potentially that this image [has] a higher ranking

[14] Gleichzeitig wird die Verwendung von DLT als Lösung für ein anderes Problem beschrieben (vgl. C2PA 2022, S. 6).
[15] Siehe hierzu Abschnitt 6.3.

> algorithmically (...) that makes it more accessible [so] more people can see it.“ (Castellanos u. a. 2022, 00:30:36)[16]

Bilder ohne CCr könnten nicht nur weniger Berücksichtigung durch Algorithmen auf Online-Plattformen finden, sondern allgemein als weniger glaubwürdig eingeschätzt werden, wie Castellanos u. a. (2022) und Feng u. a. (2023) anmerken (vgl. auch C2PA 2022, S. 6). Die zwingende Verwendung der Technologie ist seitens der C2PA zwar nicht vorgesehen (ebd.), für deren Wirksamkeit aber unabdingbar. Hieraus ergibt sich ein Konflikt, der letztlich von Anwender:innen, insbesondere Fotograf:innen, Bildredaktionen und Plattformbetreiber:innen gelöst werden soll. Gerade Menschen mit geringen technischen Ressourcen, fehlendem Zugang zu kostenpflichtiger C2PA-fähiger Software und fehlender moderner Hardware oder allgemein schwierigen Ausgangslagen könnten durch flächendeckende Implementation der C2PA einen Chancenverlust[17] erleiden. So wären es wieder „non-professional, community, non-accredited and historically marginalized communities“ (C2PA 2022, S. 2), die das Nachsehen hätten. Durch minimal lauffähige Implementationen (Im Original „minimal viable implementations“, (C2PA 2022, S. 8).), Bibliotheken in mehreren Programmiersprachen und dem *open-source* Gedanken soll dieses Risiko minimiert werden (ebd.).

Es kann nicht unerwähnt bleiben, dass sogar manipulierte Bilder mit gültigen CCr durch ein CCr-Pin auf den ersten Blick glaubwürdig erscheinen können (vgl.Versuch 3). Die im folgenden Abschnitt aufgezeigten Schwachstellen der Spezifikation und ihrer Implementationen unterstreichen den Schweregrad des Missbrauchsszenarios, welches in der Risikoanalyse unter „Manipulation“ angeführt wird. Ältere Bilder können nachträglich mit CCr versehen werden, die dann nur einen Teil der Herkunft erklären. Durch *Soft Bindings* besteht die Möglichkeit, Manifeste nachträglich zu bearbeiten, und so irreführende Informationen anzuzeigen.
Zwar lässt eine nicht repräsentative Studie, durchgeführt von Adobe-Angestellten und wissenschaftlichen Mitarbeiter:innen der University of Washington, hoffen, dass das Risiko gering bleibt (vgl. Feng u. a. 2023), dennoch bleibt es gerade in der Übergangszeit, in der noch nicht alle Webseiten und Fotograf:innen von der Technologie Gebrauch ma-

[16] Zur besseren Lesbarkeit wurden minimale Anpassungen des Interview-Transkript (in eckigen Klammern) vorgenommen, die den Sinn jedoch nicht verändern.
[17] Im Original: „opportunity loss“ (C2PA 2022, S. 2ff).

chen, und viele Bilder nachträglich mit CCr versehen werden, ein ernst zu nehmendes Risiko. Eine weiterführende Evaluation der Risiken im Hinblick auf die Erfolgschancen der Spezifikation erfolgt im nächsten Kapitel ab Seite 69.

5.4 Kritik an der C2PA und ihrer Spezifikation

5.4.1 Kritik am Entwicklungsprozess

Abdobe spielt eine zentrale Rolle bei der Erstellung und Implementierung der Spezifikaton. Es ist begrüßenswert, dass große Unternehmen gemeinsam versuchen, einen Lösungsansatz zu erarbeiten, die Dominanz mit welcher Adobe die Umsetzung forciert und bewirbt, ist jedoch problematisch. Adobe ist sowohl in der CAI als auch in der C2PA stark vertreten; Führungspositionen sind überwiegend mit Adobe-Mitarbeitenden besetzt[18]. Das führt bestenfalls zu einer effizienten Umsetzung der Spezifikation, hebelt aber mögliche Kontrollmechanismen aus.
Die erstellten Werkzeuge, Bibliotheken und die Spezifikation selbst sind *open-source*, jedoch nicht durch *open development* entstanden. Der Quellcode für das c2patool wird über Adobe-Webseiten und Adobe-kontrollierte Repositories veröffentlicht, und anschließend von Unternehmen implementiert. Der Discord Kanal für Entwickler:innen wird von Adobe-Mitarbeitenden moderiert, um mitzuwirken muss man zuvor dem Code of Conduct von Adobe zustimmen, bei unangenehmer Kritik wird man ausgeschlossen, wie es CYBERGEM passiert ist (vgl. CYBERGEM [@UltraTerm] 2024).
Der Code für Spezifikation und Software wird von einer vergleichsweise geringen Anzahl an Menschen erstellt. Das Github Repository für die Spezifikation ist öffentlich, die Themen (*issues*) werden von unterschiedlichen Usern erzeugt, *Commits* kamen laut Github Insights in den letzten 24 Monaten jedoch ausschließlich von Adobe-Mitarbeiter L. Rosenthol. Es gibt (Stand 26.03.2025) lediglich 225 User, die den Repositories der CAI folgen, was keine Mitarbeit impliziert. Das Repository für die Spezifikation wird

[18] Am präsentesten treten Andy Parsons, Santiago Lyon, Leonard Rosenthol und Pia Blumenthal sowie John Collomosso in Erscheinung. Sie sind sehr oft Co-Autoren von Case Studies, Reviews, Präsentationen, News-Beiträgen und Podiumsdiskussionen. Beispielhaft seien erwähnt: (Balan u. a. 2023), (Feng u. a. 2023), (Parsons 2023), (Collomosse u. a. 2024b) und (Bhowmik D u. a. 2024).

von 123 Usern favorisiert. Man kann schlussfolgern, dass diese geringe Anzahl an Teilhabenden[19] und Interessierten einem so großen Projekt nicht gerecht wird. Die geringe Anzahl an Mitwirkenden verstärkt ein in der Programmiergemeinschaft allseits bekanntes Problem: fehlende Diversität unter den Mitwirkenden. Zwar unterstreicht Adobe bei jeder Gelegenheit die Möglichkeit, zu partizipieren, etwa durch Podiumsdiskussionen und in `README`-Dateien, die tatsächliche Erstellung des Codes bleibt jedoch auf wenige, Adobe-Assoziierte beschränkt.

Auch der Umgang mit Sicherheitsbedenken- und Lücken wird durch Adobe gesteuert. Im Repository `c2pa-attacks` steht in den Hinweisen `CONTRIBUTING.md`[20] , dass Sicherheitsbedenken nicht öffentlich diskutiert, sondern in ein Formular eingetragen werden sollen (vgl. CAI 2023). Der Formular-Link führt auf die Seite `helpx.adobe.com`. Ähnliches wird auch im Repository `c2pa-rs` gefordert (vgl. CAI 2025b). So behält Adobe die Kontrolle über mögliche Sicherheitslücken und entscheidet intransparent wie, ob und wann die Lücken geschlossen werden.

5.4.2 Schwachstellen der Spezifikation

Die geringe Anzahl an involvierten Personen verhindert, dass Fehler und Schwachstelle der Spezifikation schnell gefunden und behoben werden können. Schwachstellen gibt es viele. Die Versuche zeigen, dass der Name im Feld *Creator* durch den Account-Namen bei Adobe und Microsoft vorgegeben ist. Da dieser frei wählbar ist, kann die Urheberschaft zumindest mittels Namen vorgetäuscht werden.
Auch zeigte sich durch die Versuche, dass das *Hard Binding* sehr schnell aufgelöst wird, auch wenn ein Bearbeitungsschritt fehlschlägt (vgl. `V6-firefly-edit1.jpg`; Abschnitt 3.5). Mit *Durable Content Credentials* soll verhindert werden, dass durch einfaches Löschen der Metadaten kein Prüfen der CCr mehr möglich ist. Dennoch bleiben gerade Wasserzeichen angreifbar - etwa durch regenerative Angriffe, bei denen Diffusion Models das Bild „nachzeichnen“, ohne das Wasserzeichen zu übernehmen;

[19] Zum Vergleich: Das core-Repository von Libre Office Writer hat 3000 *Follower* und bereits auf der ersten Seite der aktuellen *pull requests* stehen mehr verschiedene User-Namen als es Mitwirkende bei C2PA-Repositories gibt (Stand 26.03.2025).
[20] Das Dokument wurde vom Adobe-Mitarbeiter Peleus Uhley erstellt.

durch konstruktive Methoden wie Rauschfilterung, die Wasserzeichen auf Bit-Ebene zerstören können; oder durch destruktive Angriffe, bei denen starke Kompression oder Bearbeitung das Wasserzeichen derart beschädigt, dass es nicht mehr von Detektoren erkannt wird.

KI-generierte Bilder enthalten aktuell nicht automatisch CCr, auch wenn dies von den Entwickler:innen versprochen wird. OpenAI ermöglicht im Chat mit chatGPT, das generierte Bild auf Nachfrage als JPG herunterzuladen, diese Datei enthält im Gegensatz zur PNG keine CCr (Versuch 9, `V9.jpg` und `V9-png.png`). In Microsoft Designer erzeugte Bilder enthalten nur dann CCr, wenn *zuerst* das Dateiformat ausgewählt, dann die Einbindung von CCr aktiviert, und anschließend erst auf Download geklickt wurde. Der Prozess ist fehleranfällig und das Einbinden von CCr zum Zeitpunkt der Versuche optional (vgl. Versuch 8 in Abschnitt 3.3).

Aktuell gibt es nur sehr wenige Bilder im Internet, die mit C2PA-fähigen Kameras aufgenommen wurden, folglich sind die meisten Fotografien nachträglich signiert. Alle Änderungen, die vor Import in bspw. Photoshop vorgenommen wurden, werden so nicht durch die CCr nachvollziehbar. Das wird auch auf der Verify-Seite angezeigt (vgl. Abb. 3.5), eine Manipulation kann entsprechend nicht ausgeschlossen werden. So ist die Aussagekraft aller nach Aufnahme mit CCr versehenen Bilder enorm eingeschränkt.
Wie in den Versuchen 1 und 3 hinreichend bestätigt, können Metadaten, auch C2PA-Metadaten einfach gelöscht werden. Das Löschen aller Hinweise auf (vormals) existierende CCr wird durch den Einsatz von Wasserzeichen erschwert aber nicht verunmöglicht. OpenAI gibt in seinen FAQ zu: „Metadata like C2PA is not a silver bullet to address issues of provenance. It can easily be removed either accidentally or intentionally“ (OpenAI 2025).
Versuch 3 beweist, dass wichtige Informationen verschleiert oder manipuliert werden können. Eine mittlerweile behobene Schwachstelle ermöglichte es, auf binärer Ebene Metadaten zu manipulieren, ohne das es auf der Verify-Seite ersichtlich wurde. Neal Krawetz demonstrierte das im Mai 2024 live[21], indem er das auf der Verify-Seite angezeigte Datum mittels `hexedit` innerhalb von 60 Sekunden änderte, ohne dass die Verify-Seite die Manipulation anzeigte. So war es möglich, ein verifizierbares Bild mit

[21] Im Rahmen der IPTC Photo Metadata Conference 2024 entstand die Aufzeichnung mit dem Videotitel „AI and Image Authenticity“; ab Minute 20 ist zu sehen, wie Neal Krawetz die Manipulation vornimmt.

falschem Zeitstempel zu erzeugen (vgl. IPTC 2024). Versuch 6 und 3 belegen, dass sich Metadaten bearbeiten, und anschließend mit CCr versehen lassen, ganz gleich ob die Metadaten „richtige“ Informationen enthalten, oder nicht. In Krawetz’ Worten: „The thing to remember is that a strong cryptographic signature around untrusted metadata does not make the data more trustworthy“ (IPTC 2024, 00:19:17).

C2PA-fähige Kameras müssen über die notwendige Hard- und Software verfügen, um den privaten Schlüssel sicher zu speichern und das Bild signieren zu können. Es muss verhindert werden, dass mittels SD-Karte in die Kamera geschleuste Bilder ebenfalls signiert werden, obwohl sie nicht mit dieser aufgenommen wurden. Ungeklärt ist die Frage, was beim Wiederverkauf einer Kamera mit privatem Schlüssel passiert. Benutzt die zweite Person den privaten Schlüssel der ersten Person? Wird ein neues Schlüsselpaar erzeugt, und wenn ja, wie kommt es in die Kamera?

Die C2PA-Spezifikation V. 2.1 nutzt viele bereits bestehende Standards, was ihre Abhängigkeiten verstärkt: Das Format BMFF (ISO/IEC 14496-12:2022) wird für die Containerstruktur der Bilddaten genutzt, JUMBF (ISO 19566-5:2023) zur Einbettung der Metadaten, die EXIF-Spezifikation V.2.32 sowie der XMP Standard zur Definition der (kameraspezifischen) Metadaten. Das Datenformat CBOR (vgl. Abschnitt 11.1 in C2PA V2.1), diverse JPEG und andere Dateiformat-Standards dienen der Datenstruktur. Der X.509-Standard der ITU Telecommunication Standardization Sector (ITU-T) wird für das Format digitaler Zertifikate und Public Key Infrastructure (PKI)-Infrastruktur verwendet. Gerade Letzteres ist als sicherheitskritisch zu bewerten, weil das Vertrauensmodell auf eben diese Zertifikate und Vertrauensketten aufbaut.
Die von Adobe gepflegte und veröffentlichte Implementierung der C2PA-Spezifikation[22] nutzt knapp 80 Bibliotheken (Stand: 14. April 2025; (CAI 2025b)), und stützt sich überwiegend auf `openssl` für das Verifizieren der Zertifikate. `openssl` hatte in der Vergangenheit mit gravierenden Sicherheitslücken zu kämpfen[23]. In der Spezifikation heißt es: „A claim generator should use the Online Certificate Status Protocol“ (C2PA

[22] C2PA-Implementationen weden aktuell in Rust geschrieben `GitHub-contentauth/c2pa-rs`.
[23] Einen Überblick gibt die Seite `https://openssl-library.org/news/vulnerabilities/index.html`

V2.1, Abschnitt 14.5.2). Man beachte hierbei die Wortwahl „should“ [24] anstatt „shall“[25]. Würden ungültige Zertifikate von Implementierungen akzeptiert, so untergrübe dieses Verhalten das Vertrauensmodell der Spezifikation, das auf der Integrität gültiger Zertifikate und Signaturen basiert.
Einerseits führt die Einbindung gut etablierter Standards zu einer weiten Verbreitung und vereinfachten Beschreibung der Spezifikation, andererseits führen Abhängigkeiten von diesen und entsprechenden Bibliotheken zu kontinuierlichem Wartungsaufwand und vererbten Schwachstellen. Hinsichtlich der Komplexität von Systemen gibt es zwei unterschiedliche Sichtweisen: Je komplexer die Anforderungen, desto schwieriger ist es, die CCr zu fälschen, oder aber je komplexer das System, desto größer die Angriffsfläche, und dementsprechend wahrscheinlicher werden Angriffe und Manipulationsversuche. Während Ersteres den Aufwand für eine verifizierbare, aber manipulierte Bilddatei erhöht und so viele davon abhalten könnte, es zu versuchen, ist Letzteres eine Einladung an Hacker und ressourcen-starke Akteure, die Sicherheitslücken auszunutzen.

Die wohl am intensivsten diskutierte Schwäche der Spezifikation ist ihr Zertifikatsmanagement. Version 1 der Spezifikation erlaubte es, Manifeste mit selbst-signierten Zertifikaten zu unterzeichnen. Version 2.1 validiert nur noch X.509 Zertifikate, die von *Certificate Authorities* (CA) erstellt wurden, welche sich wiederum in der von C2PA verwalteten C2PA-Trust-List[26] befinden müssen. Zusätzlich gibt es noch eine Liste mit vertrauenswürdigen *Time Stamp Authorities* (TSA), die jede Implementation selbst vorhalten soll (vgl. C2PA V2.1, S. 63). Die *Trust List* ist noch sehr kurz und von Microsoft Mitarbeiter Christian Paquin im Github `c2pa-explorations`[27] veröffentlicht. Der Westdeutscher Rundfunk (WDR) und die Deutsche Welle (DW) sind bereits inkludiert, ebenso andere Langzeit-Mitglieder der CAI.
Der X.509 Standard nutzt PKI; das National Institute of Standards and Technology (USA) (NIST) merkt an, dass Zertifizierungsstellen unterschiedliche Vorgaben bei der Erstellung von Zertifikaten geben, und Opfer von Cyber-Attacken sein können (vgl.NIST 2024). Das Konzept der *Trust List* wirkt unfertig; bei so wenigen „Vertrauten“ müsste die

[24] Das Bundesamt für Sicherheit in der Informationstechnik (BSI) übersetzt das mit „sollte“ bzw. „empfohlen“ nach RFC 2119 (TR-03183 2023).
[25] Das BSI übersetzt das mit „soll“ bzw. „muss“ nach RFC 2119 (TR-03183 2023).
[26] Zum Thema *Trust list* vgl. Abschnitt 2.3.15 und 5.2.2.2.0 in C2PA V2.1
[27] Link zur *Trust List* im Repository `c2pa-explorations`: `https://github.com/christianpaquin/c2pa-explorations/tree/main/trust-lists`

Anzahl nicht-vertrauenswürdiger Zertifikate überwiegen. Wem vertraut werden kann, bestimmen also Adobe und Co, was gerade für Citizen Media problematisch werden kann. Castellanos von WITNESS erklärt im Interview „the C(2)PA, I would say, (is) not trying to create trust, but (it's) just leveraging existing relationships of trust" (Castellanos u. a. 2022, 00:08:10). Musste man bisher bei Bildern den Fotograf:innen, Bearbeiter:innen und veröffentlichenden Instanzen glauben, kommen mit CCr Algorithmen und Software hinzu, denen man vertrauen muss.

Die Summe an Negativbeispielen lässt allgemeine Zweifel aufkommen. Krawetz stellt wiederholt fest, dass veröffentlichte Bilder mitCCr Inkonsistenzen aufweisen, die an der Glaubwürdigkeit und Fälschungssicherheit dieser zweifeln lassen (vgl. Krawetz 2024c; Krawetz 2025). Seine erfolgreichen Manipulationsversuche deuten darauf hin, dass es weiterhin Sicherheitslücken gibt, und es nur eine Frage des Engagements, bzw. der Fähigkeit der Angreifer:innen ist, wie schnell und häufig Schwachstellen entdeckt und ausgenutzt werden. Es bleibt ein Katze-und-Maus-Spiel. Krawetz ist der präsenteste Kritiker der Spezifikation, und seine Hacks werden den Reaktionen zu folge auch von Adobe-Mitarbeitenden aufmerksam verfolgt. Es gibt aber auch andere Akteure, die ihre Zweifel und Kritik mehr oder weniger laut äußern: nebst dem bereits erwähnten User CYBERGEM hat auch Adam Zeloof einen Hack veröffentlicht; Andy Parsons (Adobe) reagierte mit einem Kommentar (vgl. Zeloof 2023). Insgesamt kann beobachtet werden, dass im Jahr 2023 viele Sicherheitslücken entdeckt und geschlossen wurden. Wie in Abschnitt 6.1 beschrieben wird, müssen Metadaten in sich konsistent sein, um glaubwürdig zu sein. Es braucht nur eine einzelne, nicht erklärbare Unstimmigkeit, um die Authentizität der Bilddatei als Ganzes in Frage zu stellen. Krawetz' Analysen lassen vermuten, dass es sehr schnell zu Unstimmigkeiten kommt, sofern keine perfekt C2PA-fähige Arbeitsweise eingehalten wird.

6 Synthese und Beantwortung der Forschungsfrage

Dieses Kapitel befasst sich mit der Synthese der zuvor getrennt betrachteten Gebiete, mit dem Ziel, die Forschungsfrage zu beantworten. Die Rolle von Metadaten in der Bildforensik und ihre Belastbarkeit werden in Abschnitt 6.1 betrachtet und bewertet. Es folgt in Abschnitt 6.2 die Einordnung forensischer Methoden in den Kontext journalistischer Verifikationsprozesse und eine Abschätzung potenzieller Auswirkungen der CCr auf eben diese. In Abschnitt 6.3 wird der aktuelle Stand der C2PA mit ihren Risiken und Chancen zusammengefasst, um eine Bewertung und Zukunftsprognose vorzunehmen. Abschnitt 6.4 macht deutlich, dass die Darstellung und Interpretation der C2PA-Metadaten entscheidend für den Erfolg der CAI ist. Deshalb wird ein Verbesserungsvorschlag in Abschnitt 6.5 eingebracht, der einen differenzierteren Umgang mit dem Begriff Authentizität ermöglichen soll. Zum Schluss folgen Resümee und Reflexion.

6.1 Bildforensik als Werkzeug, Metadaten als forensisches Indiz

Die Betrachtungen in Kapitel 4 verdeutlichen, dass mit Hilfe forensischer Analysen zwischen bearbeiteten und nicht-bearbeiteten Fotografien unterschieden werden kann. Auch KI-generierte Bilder können durch forensische Untersuchungen als solche identifiziert werden. Die Analyse auf Containerebene erzeugt zusätzliche Informationen, die in Verifikationsprozessen gewinnbringend eingesetzt werden können. Metadaten spielen in allen Teilbereichen der Bildforensik eine komplementäre Rolle; oft wird auf Bild- und

Pixelebene analysiert, und die entstehenden Annahmen mit den Informationen aus den Metadaten abgeglichen. Dabei ist von Vorteil, dass Bilder über „automatisch generierte Metadaten“ verfügen und dadurch „besonders informationsgesättigt verdatet“ sind (Rothöhler 2024, S. 39). Die immer realistischer werdenden Fake Videos und Bilder führen dazu, dass eine Analyse auf Bild– und Pixelebene nicht mehr ausreicht (vgl. Xiang u. a. 2021). „To handle scenarios where image content fails to explain image evolution, file metadata can be used to help fill in the gaps“ (Bharati u. a. 2019, S. 2).
Durch Analyse auf Containerebene kann nicht nur die Frage *ob*, sondern auch *wie* und ggfs. mit welchen Mitteln und welcher Teilbereich eines Bild bearbeitet wurde, erörtert werden. Somit sind Rückschlüsse auf den Entstehungsprozess möglich. Weil die meisten Metadaten automatisiert erstellt werden, Nutzer:innen oftmals nur begrenzten Zugriff darauf haben und gerade Laien oft das notwendige Wissen oder schlicht die Motivation fehlt, Metadaten aktiv zu bearbeiten, sind Manipulationen auf dieser Ebene unwahrscheinlicher, als auch Bildebene. Piva u.a. begründen dies auch damit, dass konsistentes Fälschen der Dateistruktur übermäßig kompliziert[1] ist, und entsprechende Programmierkenntnisse erfordert (vgl.Piva u. a. 2022, S. 385).
Selbst wenn keine eindeutige Beurteilung möglich ist, können Inkonsistenzen in vorhandenen Metadaten auf eine Manipulation hindeuten und weitere Untersuchungen initiieren. Ergeben sich Zweifel etwa durch erwartbare, aber fehlende oder widersprüchliche Metadaten, so muss zusammen mit Indizien aus den Bild- und Pixelanalysen evaluiert werden, ob es wahrscheinlich eine ungewollte Bearbeitung oder doch eine bewusste Manipulation gewesen ist.

Wenngleich die Möglichkeit, Metadaten zu bearbeiten oder auch zu löschen, die Belastbarkeit dieser einschränkt, bekräftigen obige Argumente[2] jedoch die Relevanz einer eingehenden Analyse. Sofern vorhanden, können Metadaten auf Plausibilität, Konsistenz und Vollständigkeit hin geprüft werden.
Auch wenn es durch fehlerhafte Einstellungen in Kamera und Computer zu falschen Zeitstempeln kommen kann, verraten Unstimmigkeiten zwischen den einzelnen Einträgen eine bewusste Manipulation dieser, oder aber eine Bearbeitung der Datei durch verschiedene Instanzen. Die C2PA-Metadaten ändern am Problem fehlender Metadaten nichts,

[1] Im Original „overly complicated“ (Piva u. a. 2022, S. 385).
[2] Vergleiche hierzu auch das Zitat auf Seite 43 von Piva u. a. (2022).

denn auch sie können gelöscht werden (siehe Abschnitt 3.1). Auch werden Metadaten durch C2PA nicht glaubhafter, denn sie können zuvor manipuliert worden sein (siehe Abschnitt 3.7). Durch KI-Einsatz erzeugte C2PA-Einträge können einfach gelöscht, und das Bild nachträglich wieder mit CCr versehen werden, sodass kein Rückschluss auf die verwendete KI möglich ist (siehe Abschnitt 3.6). Viel hilfreicher für Forensiker:innen und Journalist:innen wäre es, wenn Bildbearbeitungssoftware und Online-Plattformen grundsätzlich alle Metadaten erhalten würden, sodass anhand der Gesamtheit aller Metadaten der Entstehungsprozess eines Bildes möglichst vollständig nachgewiesen werden kann. Eine Zusatzoption im Exportieren-Dialog etwa bei Photoshop könnte „alle ursprünglichen Metadaten beibehalten“ und so einen Export mit ausreichend *provenance* Informationen ermöglichen, welche mindestens genauso vertrauenswürdig wären, wie ein validiertes C2PA-Manifest.

Das Beibehalten aller Metadaten kann auch bei der Rekonstruktion von Kriegsverbrechen helfen: Das Team des Syrian Archives nutzte die Zeitstempel veröffentlichter Bilder und Videos, um Geschehnisse zeitlich einzuordnen. Außerdem wurden Waffen-, oder Giftart in die Metadaten geschrieben (vgl. Archive 2016). Die Citizen Journalists von Belling Cat beschreiben regelmäßig in ihren Recherchen, wie sie Bildinhalte und enthaltene Metadaten analysieren, um die Verbreitung von Falschinformationen und verschleiertes Kriegsgeschehen aufzudecken: Das russische Verteidigungsministerium hatte die Zeitstempel mehrere Satellitenaufnahmen gefälscht, um den Vorwurf zu zerstreuen, das russische Millitär sei für den Abschuss des Passagierflugzeugs (MH17) verantwortlich (es berichtete Dillon (2025) von der DW). Die Folterungen im Gefängnis Abu Ghraib wurden durch die Veröffentlichung von Bild- und Videomaterial an die Öffentlichkeit gebracht; durch Metadaten konnte ein zeitlicher Verlauf erstellt und Opfer sowie Täter teilweise identifiziert werden (es berichtete unter anderem Walsh 2006 im SPIEGEL). Diese Beispiele zeigen die zentrale Bedeutung von Bild- und Videomaterial inklusive Metadaten in verschiedenen Kontexten.

Die Fülle an Metadaten in Bildern, ihre Informationsdichte und die Belastbarkeit bei Vollständigkeit sprechen für die Verwendung von Metadaten als Indiz. Sogar fehlende Metadaten oder Inkonsistenzen können ein Indiz sein und Falschinformationen aufdecken, wie das MH17-Beispiel verdeutlicht. Bildforensische Methoden inklusive Analyse enthaltener Metadaten sollten weiterhin und mit C2PA verstärkt Einzug in Authentifizierungs-

und Verifikationsprozesse erhalten. Zwar beseitigt die C2PA einschränkende Argumente nicht, erzeugt aber zusätzliche Informationen, die für containerbasierte Methoden und Analysen genutzt werden können.

6.2 Journalistische Verifikationsprozesse und Einfluss der C2PA

Die Bild-Authentifizierung ist Teil des Verifikationsprozesses, wenn Bilder veröffentlicht werden. Bei der Verifikationen müssen Kontextinformationen mit einbezogen werden. Wer hat die Fotografie wann, wo aufgenommen? Was zeigt das Bild und stimmen angegebenen Informationen mit den Aussagen des Bildes und der Metadaten überein? Ist die Quelle glaubwürdig und kann das Bild einer objektiven Überprüfung standhalten (vgl. Stern 2020, S. 120ff)? In gewisser Weise lässt sich so bei der Verifikation auch von einem Plausibilitätscheck sprechen: Dateiinhalt und Kontextinformationen müssen übereinstimmen; zusammen mit den Umständen der Veröffentlichung kann dann beurteilt werden, ob die Behauptung plausibel ist. Es gilt jedoch: auch wenn die Aussage auf Bildebene plausibel erscheint, muss auf den anderen Ebenen eine Manipulation ausgeschlossen sein, um das Bild als glaubwürdig einstufen zu können. Gerade bei Bildern aus unbekannter Quelle rät die SWGDE (2025), zu prüfen, wer Zugriff auf das Bild hat, über welche Fähigkeiten, Hard- und Software diese Personen verfügen und ob eine Motivation und genügend Zeit für eine Manipulation vorhanden war. Weiterhin muss genau überprüft werden, ob es sich um eine Rekontextualisierung handelt (vgl. Welchering 2020, S. 33), etwa wenn das Bild zwar authentisch ist, aber nicht das zeigt, was im dazugehörigen Text oder Titel angegeben wird. In jüngerer Vergangenheit werden insbesondere eben diese Rekontextualisierungen vom Deutschen Presserat missbilligt[3]. Im Fall mit dem Aktenzeichen 0179/22/2 ging es beispielsweise um ein fünf Jahre altes Video bzw. Standbild, dass im Zusammenhang mit einem aktuellen Krieg veröffentlicht wurde.

War die Bildlandschaft in Zeitungen Ende des zwanzigsten Jahrhunderts noch überwie-

[3] Wegen Einsicht der beschuldigten Redaktion erfolgte im erwähnten Fall lediglich ein Hinweis durch den Presserat.

gend von Berufsfotograf:innen geprägt, und die Glaubwürdigkeit ihrer Bilder an das Vertrauen zwischen Redaktion und Fotograf:in geknüpft, so erhalten Stock-Footage, Smartphone-Bilder und synthetische Inhalte immer mehr Einzug in die mediale Berichterstattung, wodurch das bisherige Vertrauensmodell ins Wanken gerät. Während Loosen u. a. (2020) beobachten, dass weiterhin Intuition und manuelle Recherche die journalistische Praxis dominieren, beschreibt Gerling (2022), dass bildforensische Methoden wichtiger werden. Das klassische Handwerkszeug des Zwei-Quellen-Prinzips[4] zusammen mit Werkzeugen aus dem Bereich der OSINT Techniken bilden eine gute Basis bei der Verifikation von Bildern, wenngleich synthetische Bilder damit nicht eindeutig erkannt werden können. Bilderrückwärtssuche, Browserbasierte Software wie forensically oder das Chrome-Plugin „InVid“ sowie Datenbanken und Webseiten aus dem OSINT Werkzeugkasten der Organisation Bellingcat werden vielfach beschrieben, wenn es um die Verifikation von Informationen und Bildern geht (vgl. Osing 2022; Welchering 2020; Gerling 2022). Die Analyse der Metadaten ist dabei grundlegend (vgl. Welchering 2020). Am intensivsten werden die genannten Techniken in den sich entwickelnden Faktencheck-Einheiten angewandt; man kann die Entstehung dieser neuen Organisationseinheiten auch als Anpassungsleistung der Branche an die wachsende Herausforderung durch KI und Social Media bzw. Citizen Media begreifen.

Bilder, welche zur Vermittlung von „Wahrheit“, als „Produkte von Augenzeugenschaft“ (Leifert 2007, S. 248) quasi als „Beweis“ (Osing 2022, S. 179) in der Berichterstattung eingesetzt werden, müssen trotz etwaiger Bearbeitung authentisch sein. Der Deutsche Presserat unterscheidet zwischen dokumentarischen Abbildungen, Illustrationen, Montagen und Symbolbildern. Erstere müssen wahrheitsgetreu wiedergegeben werden, also authentisch sein. Letztere, gerade wenn sie bei flüchtigem Blick dokumentarisch wirken, müssen „deutlich wahrnehmbar in Bildlegende bzw. Bezugstext als solche erkennbar“ sein (Deutscher Presserat 2025, Ziffer 2). Das gilt auch für KI-generierte Bilder, wie die jüngste Spruchpraxis zeigt: Im Beschwerdefall 0770/23/1-BA wurde missbilligt, dass das Portrait der KI-Reporterin nicht direkt als KI-generiertes Symbolbild gekennzeichnet war.

Journalist:innen werden mehr und mehr forensische Methoden anwenden müssen, um Bilder zu authentifizieren und Informationen zu verifizieren. Sie würden von funktionie-

[4] Vgl. hierzu etwa Osing 2022, S. 179.

renden, lückenlosen Implementierungen der C2PA-Spezifikation profitieren, weil auch technisch weniger versierte Menschen mit Content Credentials ein Werkzeug erhielten, mit dem sich Metadaten als Informationsquelle im Verifikationsprozess nutzen ließen. In jedem Fall sollten Journalist:innen zugelieferte Bilder kritisch auf Manipulation hin untersuchen. Wenngleich Metadaten bearbeitet sein können, so kann eine Analyse unentdeckte Manipulationen auf Bildebene aufdecken. Interessant ist der Gedanke von Loosen und Solbach, dass ein automatisierter Faktencheck, der durch Eingaben der Nutzer:innen gesteuert werden kann, unempfindlicher gegenüber dem Vorwurf der „tendenziösen Vorauswahl“ wäre (Loosen u. a. 2020, S. 186). Hierzu müssten Algorithmen mit Metadaten, dem Bild selbst und Kontextinformationen sowie Nachrichtentext und Eingaben der Nutzer:innen gefüttert werden.

6.3 C2PA - Auf dem Weg, ein Teil der Lösung zu sein

Die *Chain of Custody*, der typische Lebenszyklus eines digitalen Bildes, besteht aus den Aktionen erstellen, verwalten, bearbeiten, exportieren und veröffentlichen.
Hardware-Hersteller wie Leica, Sony, Canon, Fuji und Nikon haben Kameras im Sortiment, die C2PA-Manifeste erzeugen und noch in der Kamera signieren (Stand April 2025). Sony hat für einige Modelle angekündigt, die Funktion über ein Firmwareupdate zur Verfügung zu stellen. Für Smartphones hat das Team des Guardian Projects die App ProofMode entwickelt. Sie erzeugt beim Teilen eines mit ihr aufgenommenen Fotos einen Zip-Ordner, der sowohl die Bilddatei, als auch Tabellen-Dokumente, Schlüssel-Dateien sowie eine JSON-Datei enthalten. Das Verifizieren ist nur über `https://proofmode.org/verify` mit Hilfe des ZIP-Ordners möglich. Das PGP-Zeritifikat wird bei App-Installation erzeugt, es werden keine persönlichen Daten erhoben, aber (optional) Gerätedaten in den Proof-Dateien gespeichert. Wird ein bereits existierendes Bild in der App aufgerufen und signiert[5], ist das im Eintrag `Notes=ProofMode v2.5.0-RC-1 autogenerated=false` ersichtlich. Die Implementation ist in sich schlüssig und wird bereits von Menschenrechts-Aktivist:innen verwendet, ist aber auf Grund der stark be-

[5] Für ein Beispiel sei auf `V2_2.jpg` verwiesen.

grenzten Kameraeinstellungen keine massentaugliche Lösung.
Qualcom kündigte im Oktober 2023 an, gemeinsam mit Truepic die Snapdragon 8 Mobilplattform der dritten Generation C2PA-fähig zu machen (vgl. Parsons 2024a). Damit würden auch entsprechende Smartphones in der Lage sein, mit PKI kryptografische Signaturen und damit verifizierbare C2PA-Manifeste in Bilder einzubetten.
Bildverwaltungssoftware ist gerade bei professionellen Fotograf:innen die erste Software-Instanz, in der die Bilder verarbeitet werden. Camera Bits Photo Mechanic® hat die Spezifikation implementiert (vlg. Orlosky 2024), weitere Softwarehersteller wie Photoshelter (vgl. PhotoShelter 2024) und Fotoware (vgl. Stephan 2024) sind Mitglied der C2PA, werben jedoch nicht offensiv mit eigenen Implementationen.

Am anderen Ende der Bildverarbeitungskette stehen Plattformen, Internetseiten und Online-Verlage.
Das Technologieunternehmen Cloudflare bietet seit Beginn 2025 die Möglichkeit, auf Cloudflare Image hochgeladene Bilder mit CCr zu versehen oder bereits vorhandene CCr beizubehalten. Cloudflare ist laut Gray (2025) für 20 % des Internetverkehrs verantwortlich; in seinem Online-Artikel wird Andy Parsons (Adobe) zitiert: „Cloudflare Images's implementation will ensure the last-mile delivery of Content Credentials to the end user when a site owner or content creator opts to preserve them“ (Gray 2025), das Beibehalten von CCr ist folglich optional.
Social-Media-Plattformen haben die Spezifikation zumindest auf deutschen Seiten bisher nicht umgesetzt (vgl. Abschnitt 3.2). Auch LinkedIn, die Plattformen des Meta-Konzerns und TikTok geben an[6], Teil der C2PA zu sein und CCr integrieren zu wollen. Der Wille scheint vorhanden, die Umsetzung lässt jedoch auf sich warten.
Google will die C2PA-Manifeste nutzen, um das Feature „Infos zu diesem Bild“ anzureichern. `V4-edit3.jpg` enthält gültige CCr, das Ergebnis in Abb. 6.1 zeigt jedoch noch keine Hinweise auf die Manipulation mittels KI; auch mit chatGPT generierte Bilder werden von Google noch nicht als synthetisch erstellt ausgewiesen (Stand 03. Mai 2025).
Die Videoplattform YouTube will enthaltene C2PA-Metadaten nutzen, um in der Beschreibung einen Hinweis auf den Ursprung des Videos anzuzeigen. In den Richtlinien steht: „Damit die Offenlegung 'Mit einer Kamera aufgenommen' in der erweiterten Videobeschreibung angezeigt wird, müssen Creator beim Aufnehmen ihrer Videos Tools

[6] Vgl. hierzu Corrigan 2024; Clegg 2024; TikTok 2024

Abbildung 6.1: Google's „Infos zu diesem Bild" zu `V4-edit3.jpg` zeigt keinen Hinweis, dass es mit KI bearbeitet wurde.

mit integrierter C2PA-Unterstützung (Version 2.1 oder höher) verwenden" (YouTube 2025). Die Plattform übernimmt folglich die Interpretation der C2PA-Manifeste. Das erste mit CCr versehene und auf YouTube mit entsprechendem Hinweis veröffentlichte Video[7] wurde mit der Truepic Capture Camera App aufgenommen.
Eine Auswahl von Verlagen, die C2PA bereits implementiert haben, wurde bereits in Abschnitt 2.2.1 vorgestellt.

Die Darstellung der C2PA-Informationen auf Verify-Seiten ist für Rezipient:innen entscheidend, und somit für den Erfolg der CAI maßgeblich. Aktuell ist die Benutzbarkeit der CAI-Verify-Seite[8] verbesserungswürdig (vgl. Kapitel 3 und Abschnitt 5.4.2). Microsoft hat seinen Verify-Service vor Beginn der Versuche deaktiviert; die Seite[9] der IPTC findet man nur zusammen mit dem Stichwort IPTC, weshalb davon auszugehen ist, dass überwiegend Journalist:innen darauf zurückgreifen werden. Eine gänzlich von involvierten Instanzen unabhängige, frei zugängliche Möglichkeit zur Überprüfung von CCr gibt es derzeit nicht. Das Exiftool kann C2PA-Metadaten lediglich anzeigen; so lässt sich immerhin prüfen, ob ein Bild C2PA-Metadaten enthält. Das c2patool ist ein Kom-

[7] Titel des Videos: I am really at the zoo
[8] Die CAI-Verify-Seite ist erreichbar unter `https://contentcredentials.org/verify`.
[9] Die IPTC-Verify-Seite ist unter `https://originverify.iptc.org/` erreichbar.

mandozeilenprogramm, entsprechend nicht für Rezipient:innen gedacht. Es ist jedoch hilfreich, wenn die auf den Verify-Seiten gezeigten Informationen nicht ausreichen.

Zwar erfolgte hier lediglich eine kleine Auswahl aus der großen Menge an CAI-und C2PA-Mitgliedern und deren Implementationen, dennoch kann es verwundern, dass trotz der namhaften Unterstützer:innen so wenige CCr-Pins im Internet zu sehen sind. Viele der auffindbaren Beispiele enthalten Fragezeichen, wie Neal Krawetz in seinem Blog regelmäßig beleuchtet (vgl. etwa Krawetz 2024b; Krawetz 2024c). Mögliche Gründe für die geringe Anzahl von Bildern mit CCr sind:

- Nur wenige Menschen haben Zugriff auf eine C2PA-kompatible Kamera.
- C2PA-fähige Bildverwaltungssoftware ist kostenpflichtig.
- Das Erstellen von Content Credentials bei der Bildbearbeitung in Adobe Software ist oft noch optional.
- Die Mehrzahl hochgeladener Bilder auf Social-Media-Plattformen ist mit Smartphonekameras ohne C2PA-Technologie erstellt und mit Apps ohne C2PAImplementation bearbeitet worden.
- Internetseiten zeigen das CCr-Pin nicht automatisch an, es erfordert derzeit eine aktive Installation von Erweiterungen.

Insgesamt ist zu beobachten, dass Anfang und Ende der *Chain of Custody* unzureichend mit C2PA-fähigen Lösungen ausgestattet sind. Deshalb lahmt die systemweite Umsetzung.
Möglicherweise schrecken Unternehmen auch auf Grund der vorhandenen Risiken, hohem Implementationsaufwand und ungeklärten Fragen im Hinblick auf die systemweite Anwendung der Spezifikation vor einer eigenen Integration der Technologie zurück.

Die eigenen Versuche und Ausarbeitungen anderer Autor:innen haben Schwachstellen offenbart und berechtigte Risiken aufgezeigt. Versuch 4 bestätigt, dass bereits wenige Bearbeitungsschritte in Photoshop die Dateigröße signifikant erhöhen können: `V4-edit3.jpg` hat eine um 21% erhöhte Dateigröße im Vergleich zu `V4-edit2.jpg`. Die Auflösung und Kompressionsrate blieben gleich, folglich müssen nicht sichtbare

Daten[10] hinzugekommen sein. Rathi u. a. (2024) unterstützen diese Beobachtung durch eigene Untersuchungen: Sie konnten zeigen, dass gerade beim viel verwendeten jpg-Format die Daten von nur einem Manifest mehr Speicherplatz verbrauchen, als die Bilddaten selbst, bei drei Manifesten ist die Datei mit CCr mehr als doppelt so groß (vgl. Abb. 6.2). Sie schlussfolgern: „[T]he scalability design goal could stand as one of the

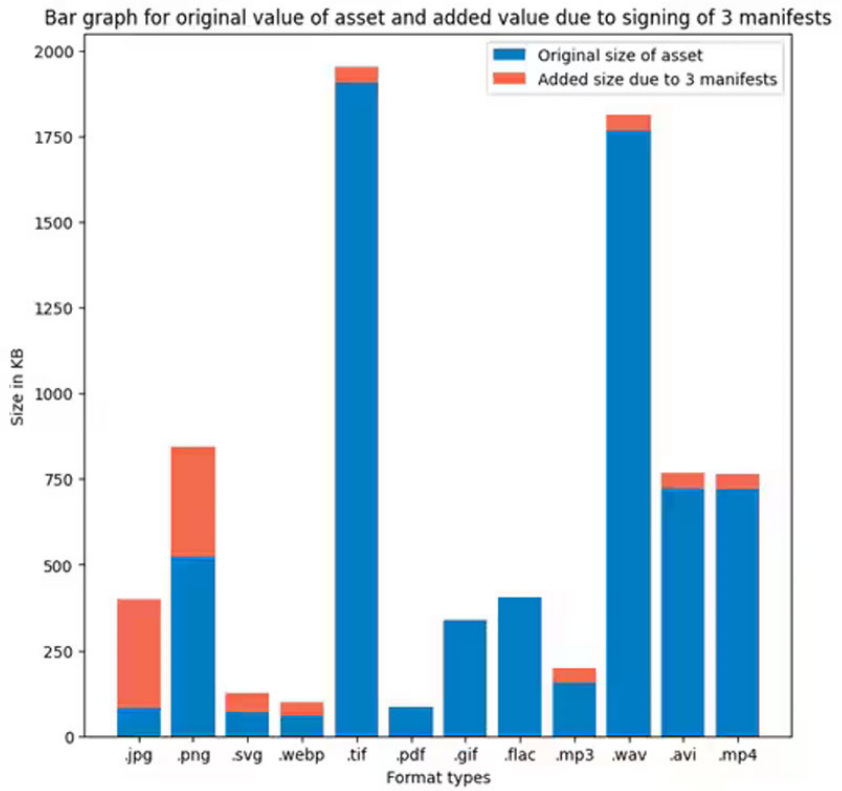


Abbildung 6.2: Grafik aus (Rathi u. a. 2024); sie zeigt die Dateigröße mit und ohne Manifeste für verschiedene Containerformate.

biggest challenges to be overcome in future“ (Rathi u. a. 2024). C2PA begrenzt die Größe der integrierten Daten nicht, was die Thematik weiter verschärft und Cyberattacken durch Einbindung von schädlichem Code oder aber Entnahme von persönlichen Daten ermöglicht (ebd.).

[10] Zur Überprüfung der Prozentzahl wurde `V4-edit3.jpg` kopiert und alle Metadaten mit Exiftool gelöscht: Die Zieldatei `NV4-edit4.jpg` ist um 25% kleiner als die Version mit einem Manifest. Die Differenz von 4% ist auf bereits zuvor vorhandene Metadaten zurückzuführen.

Auch die in Abschnitt 5.3 ausführlich beschriebene Gefahr der epistemischen Ungerechtigkeit ist nicht zu vergessen. Eine breitere Einbindung von betroffenen Gruppen in den Entwicklungsprozess, groß angelegte Bildungskampagnen und frei zugängliche Software für alle Abschnitte des Lebenszyklus eines Bildes müssen unterstützt werden, um dieses Risiko zu minimieren.
Um die Deutungsverschiebung des Konzeptes „Authentizität“ durch die C2PA abzufedern, bedarf es umfangreicher Kommunikation mit starkem Fokus auf Aufklärung und Technik-Ermächtigung. Es bleibt weiterhin fraglich, unter welchen Umständen ein KI-generiertes Bild „authentisch“ genannt werden sollte. Eine breitere gesellschaftliche Debatte dazu steht noch aus.

Trotz der in Abschnitt 5.3, Abschnitt 5.4 und zuvor genannten Schwächen darf das Potenzial der Spezifikation nicht übersehen werden, denn sie befindet sich in Arbeit und Schwachstellen können noch behoben werden. CCr ermöglichen es, den Ursprung von Bildern zu ergründen, was vor allem für Medienschaffende und Verlage bei der Auswahl glaubwürdiger Inhalte hilfreich sein kann. Authentische Bilder sind wichtig, denn sie prägen Gesellschaften, Demokratien und Geschichte. Content Credentials können außerdem dabei helfen, synthetische von echten Bildern zu unterscheiden.
Besonders das Ziel der C2PA, den Nutzen der Technologie nicht nur technisch versierten, sondern allen Menschen zugänglich zu machen, ist bemerkenswert. Denn am Ende entscheiden die Rezipient:innen, ob sie einer Bildaussage glauben, oder sie hinterfragen. Die breite Umsetzung der C2PA-Spezifikation kann Fake News nicht verhindern; sie kann jedoch Aufmerksamkeit erzeugen und Menschen befähigen, selbst aktiv zu werden und informierte Entscheidungen zu treffen. Dafür müssen technische Lösungen stets so gestaltet werden, dass alle Nutzer:innen profitieren.

Insgesamt ist die Umsetzung der Spezifikation als sinnvoll einzustufen, sofern Implementationen sicher sind und ihre Nutzung verständlich vermittelt wird. Die einseitige Steuerung durch Adobe ist kritisch zu bewerten (vgl. S. 55), könnte aber durch unabhängige Audits ausgeglichen werden. Auch wenn Open-Source-Projekte mehr Zeit benötigen, hängt der Erfolg der Spezifikation von einer breiten Akzeptanz ab, sodass die Einbindung von Open-Source-Gemeinschaften, mehr Studien zur Wahrnehmung von CCr und externe Sicherheitsaudits das Vertrauen in die Spezifikation stärken können.

6.4 Rezipient:innen als entscheidender Faktor

Erklärtes Ziel der CAI ist es, jedem Menschen zu ermöglichen, die C2PA-Metadaten nutzen zu können, um die Authentizität eines medialen Inhaltes zu beurteilen (vgl. Abschnitt 2.2.1). Aktuell findet sich im besten Fall ein kleiner CCr-Pin am oberen rechten Rand des Bildes, der das Vorhandensein von Herkunftsnachweisen andeutet. Ein Positivbeispiel ist die Webseite des Arizona Secretary Of State Office (*AzSoS Certified Media* 2025).

Zwar verlangen die CCr mehr Eigenleistung seitens der Rezipient:innen, ermächtigen sie aber gleichzeitig, sich eine eigenen Meinung zu bilden (vgl. S. 52). Sowohl die CAI als auch Beobachter:innen merken an, dass für die erfolgreiche Nutzung der Technologie die Aufklärung der gesamten Bevölkerung entscheidend ist (vgl. Feng u. a. 2023, S. 4; NSA u.a. 2025, S. 8). Rezipient:innen müssen wissen, wie die Technologie funktioniert, was man aus CCr an Erkenntnisgewinn erlangen kann, und welche Informationen nicht durch sie verifiziert werden können. Schnittstelle zwischen Technik-Dienstleister:in und Rezipient:in ist eine Verify-Seite.

Die für nicht technikaffine Menschen konzipierte CAI-Verify-Seite zeigt derzeit eine begrenzte Menge an Informationen. Dennoch gibt es viele Begriffe und Namen, insbesondere in den Manifesten selbst, deren Bedeutung vielen fremd sein dürfte: Was ist „Microsoft Content Integrity Web Application 1.0“? Was genau bedeutet die Aktion „importiert“ und wer hat denn nun das Foto geschossen? Wofür steht `trainedMediaAlgorithm` und was hat es mit der `JSON:CredentialSubjectID` auf sich? Selbst wenn man mit den Begriffen arbeiten kann, und die unbekannten Namen recherchiert hat, bleibt die Frage: Ist das Bild nun authentisch? Wo wurden Änderungen vorgenommen? Wann[11] wurde fotografiert oder generiert, oder beides? Derzeit kann nicht zwischen teilweise, oder ganz KI-generierten Bilder unterschieden werden (vgl. S. 28). Will man ein manipuliertes Bild überprüfen, erscheint ein Hinweis wie in Abbildung 3.7; was genau manipuliert wurde, ist dabei nicht ersichtlich.

Für Kreative ist auch wichtig, dass Vorschau und Veröffentlichung übereinstimmen. Ein Vergleich von Abb. 3.5 (S. 28) mit Abb. 3.10 (S. 37) legt offen, dass die Vorschau in PS akkurat ist.

[11] Das Verify-Ergebnis zeigt lediglich, wann das Manifest zuletzt signiert wurde.

Bisher gibt es nur wenige empirische Studien, die untersuchen, welchen Einfluss *provenance* Informationen auf Rezipient:innen haben. Feng u. a. (2023) haben unter Aufsicht oder Mitwirken von Adobe-Mitarbeitenden Pia Blumenthal und Andy Parsons ein Online-Experiment mit 595 englisch-sprachigen Teilnehmenden durchgeführt. Sie kommen zu der Schlussfolgerung, dass die Einbindung von mehr als drei Manifesten in die dargestellten Informationen auf Nutzer:innen verwirrend wirkt (vgl. Feng u. a. 2023, S. 11). Unvollständige oder nicht vorhandene Informationen erzeugen eine größere Skepsis, als vollständige Informationen (vgl. Feng u. a. 2023, S. 5). Es ist davon auszugehen, dass Bilder ohne CCr weniger glaubwürdig erscheinen werden, ganz gleich ob sie authentisch sind oder nicht[12]. Es hilft nicht, dass Beteiligte unentwegt betonen: „The presence or absence of provenance information should not be seen as an automatic endorsement that a post is 'true' or 'false.' [sic!]“ (*The News Provenance Project* 2025). Die genannte Studie zeigte den Teilnehmenden neun Social-Media-Beiträge, davon waren vier manipuliert oder irreführend[13] und untersuchte verschiedene Variablen und deren Einfluss auf die Glaubwürdigkeit der Bilder und der Gesamtaussage. Bei bearbeiteten oder manipulierten Bildern haben *provenance* Informationen zu einer größeren Skepsis geführt. Bei unbearbeiteten Bildern konnte (zumindest bei nationalen Nachrichtenbeiträgen) keine signifikante Auswirkung der zusätzlichen Informationen festgestellt werden.

Insgesamt zeigt sich, dass zwischen der Menge gezeigter Informationen und der Intensität vorgenommener Interpretation durch die Verify-Seite abgewägt werden muss. Vorteilhaft wären verschiedene Detailgrade und Interpretationsstufen, die kontextbezogen und transparent angeboten werden sollten, um Rezipient:innen das schrittweise Erlernen und Verstehen zu ermöglichen. Zuvor ist jedoch in einer breiten, öffentlichen Debatte festzulegen, wann ein Bild als authentisch bezeichnet werden kann und wann nicht.

[12] Vgl. hierzu NSA u.a. 2025, S. 6 und das Zitat auf Seite 53.

[13] Irreführend hieß im Kontext der Studie nicht zwingend, dass die Bilder bearbeitet waren, sondern lediglich, dass sie nicht zur Beitrags-Aussage passten (*composited, claim disagree*).

6.5 Authentizitäts-Skala - kontextbezogene Interpretation der CCr ermöglichen

In Abschnitt 5.1 ist bereits deutlich geworden, dass Authentizität kein einfaches Konstrukt ist. Die Wahrnehmung der Spezifikation durch genannte Autor:innen in Abschnitt 5.1.2 deutet darauf hin, dass Authentifizierung im Kontext der C2PA vornehmlich als Verifikation des Entstehungsprozesses eines Bildes betrachtet wird. Verifizierte Herkunftsinformationen haben jedoch keine unmittelbare Aussagekraft im Hinblick auf die Authentizität und Verifizierbarkeit der Bildaussage. Eine Bewertung nur auf Grundlage von Metadaten ist selten ausreichend. Wenn die C2PA bei validierten CCr generalisiert von einem authentischen Bild spricht, verkennt sie ethische Kriterien im Journalismus, entkoppelt den Begriff der Bildauthentizität von der allgemeinen Umschreibung[14] „so und nicht anders ist es gewesen“ und ermöglicht die legitimierte Verwendung KI-generierter Bilder im Nachrichtenkontext. Es regt zum Nachdenken an, wenn Truepic von einem authentifizierten deepfake Video spricht (vgl. truepic 2025).
Die Unterscheidung zwischen authentischen und nicht-authentischen Bildern wird mit C2PA auf analytischer Ebene zunehmend erschwert. Zwar werden KI-generierte Bilder durch CCr als solche gekennzeichnet, das ist jedoch erst auf den zweiten Blick ersichtlich. Das Icon, welches auf Bildern mit CCr erscheinen soll, muss eine genauere Differenzierung ermöglichen, schon auf den ersten Blick. Bei der Ausweitung des Konzepts auf Audio- und Textdateien muss außerdem darüber nachgedacht werden, inwiefern Unterschiede zur Bildauthentizität bestehen. Auch im künstlerischen Bereich wird Authentizität anders gewertet, wodurch die Verwendung von CCr anders kommuniziert werden muss und auch andere Ziele verfolgen kann.

Im Nachrichtenkontext entscheidet das Ausmaß der Bearbeitung zwischen authentischen und nicht-authentischen, oder gar manipulierten Bildern. Die Grenze zwischen erlaubter Bearbeitung, durch die ein Bild in seiner Authentizität nicht verletzt wird, und unerlaubter Bearbeitung ist zusätzlich kontextabhängig (vgl. Krämer u. a. 2018, S. 12).
Im journalistischen Kontext sind Bearbeitungen erlaubt, welche die technische Qualität des Fotos verbessern, nach dem Motto „Alles, was früher im Labor ohne Schere ging, ist

[14] Vgl. Krämer u. a. 2018 im Titel; Leifert 2007, S. 247.

(...) gestattet“ (Rossig 2014, S. 211). Dazu zählen bei der Associated Press (AP) unter anderem Zuschnitte, Rauschreduktion und farbliche Anpassungen (vgl. Associated Press 2025).
Unerlaubte Bearbeitungen sind solche, welche die Bildaussage verändern. Man spricht dann von Manipulation, insbesondere wenn die Fälschung mit Vorsatz[15] vorgenommen wurde. Leifert (2007) und Gerth (2018) zählen verschiedene Arten von Manipulation auf: Löschen von Informationen, Einfügen von Informationen, Fotomontagen, falsche Beschriftung (Kontextfälschung), Inszenierung, Simulation künstlicher Welten und Veränderung von Bildparametern. Letzteres ist streitbar, wird jedoch mit einem bekannten Beispiel begründet, bei dem allein durch farbliche Anpassungen eine signifikante Änderung der Bilddaussage vorgenommen wurde (vgl. Gerth 2018, S. 15). Adobe-Assoziierte sprechen auch bei redaktionellen Änderungen von „manipulation for editorial purposes“ (Collomosse u. a. 2024a), was die Grenze zwischen Bearbeitung und Manipulation verschwimmen lässt.

Aus den unterschiedlichen Auffassungen zum Begriff der Authentizität und der zum Teil unscharfen Trennung zwischen erlaubter Bearbeitung und Manipulation entsteht die Frage, wie - gerade im Hinblick auf aktuelle Herausforderungen durch KI-Einsatz - mit dem Konstrukt beim globalen Unterfangen der Content Credentials umgegangen werden soll. Sowohl Journalist:innen, als auch Rezipient:innen sollen in der Lage sein, mittels CCr die Authentizität eines Bildes zu beurteilen. Nun könnte man sagen, weil die Spezifikation lediglich Hinweise geben will, könne ja jeder und jede bei seiner bzw. ihrer individuellen Auffassung bleiben. Allerdings wird wie in Abschnitt 5.1 dargelegt, durch die praktische Umsetzung und Kommunikation der C2PA eine von der journalistischen und geisteswissenschaftlichen Auffassung abweichende Verwendung des Begriffs vorgenommen. Es bedarf einer differenzierteren Versprachlichung und bestenfalls auch Visualisierung[16] des Konstruktes Authentizität, um dem Medium Bild als Übermittler von Wahrheit gerecht zu werden.

CAI und C2PA nutzen den Begriff Authentizität ekzessiv und unterscheiden nur zwischen authentisch und fake. Ein derart simpler Ansatz wird der allgemeinen Komplexität

[15] Wann genau eine Täuschungsabsicht vorliegt, ist nach Analyse der Spruchpraxis des deutschen Presserates durch Leifert (2007, S. 220ff) nicht eindeutig festzustellen.
[16] Man könnte etwa in Form unterschiedlich farbiger CCr-Icons eine Differenzierung vornehmen, wie der Vorschlag in diesem Abschnitt nahelegt.

jedoch keinesfalls gerecht. Parsons möchte die CCr ähnlich leicht verstanden wissen, wie der NutriScore[17]; das wird von vielen Instanzen aufgegriffen (vgl. NSA u.a. 2025; Ryan-Mosley 2023; Collomosse u. a. 2024a). Die IPTC liefert durch den Metadaten-Eintrag `digitalSourceType` bereits einen differenzierteren Ansatz, indem sie zwischen geringer und starker, sowie zwischen menschlicher und algorithmischer Bearbeitung unterscheidet (vgl. Abschnitt 5.2. Kombiniert man Parsons Wunsch und die Vorlage der IPTC, so könnte man Kriterien an C2PA-Metadaten knüpfen und automatisiert eine differenziertere Kategorisierung der Bilder vornehmen und anschließend visualisieren. Abbildung 6.3 bietet hierfür eine Diskussionsgrundlage. Die Y-Achse illustriert den

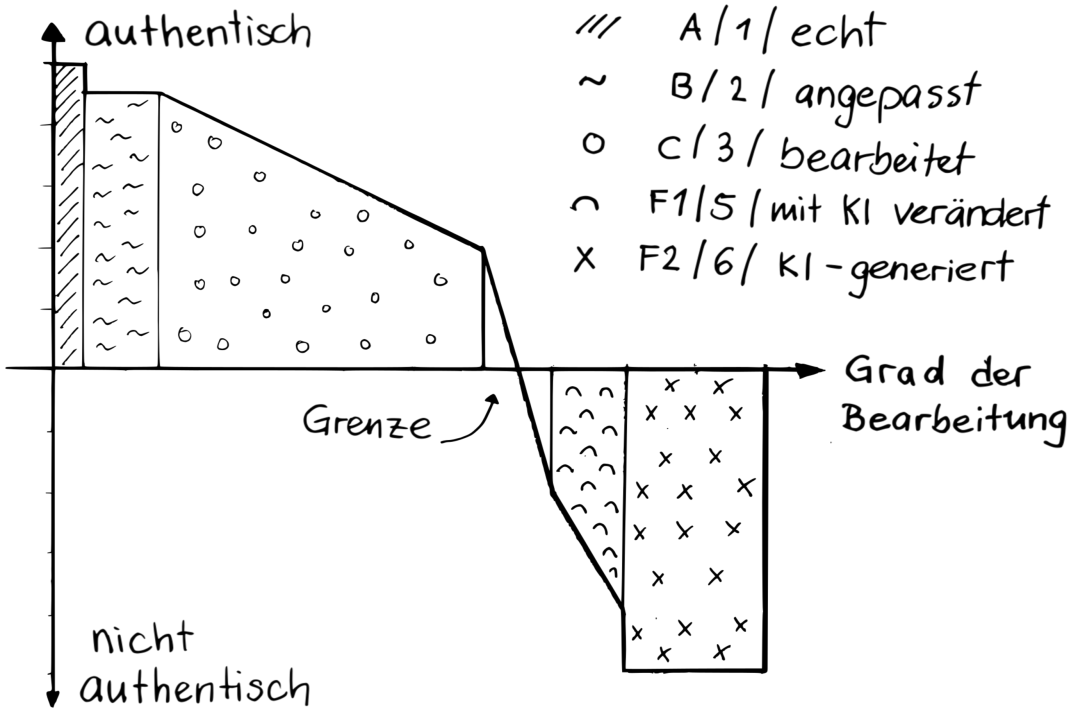


Abbildung 6.3: Authentizität in Abhängigkeit vom Grad der Bearbeitung, eigene Darstellung

Grad der Authentizität, wobei Bilder im negativen Bereich als nicht-authentisch anzusehen sind. Die X-Achse zeigt den Grad der Bearbeitung. Die genaue Position der Grenze

[17] Kritik am NutriScore sei hierbei ignoriert, es geht um das Prinzip, eine Abstufung auf Grund bestimmter Kriterien vorzunehmen, die dann auf einfache Weise visualisiert werden kann.

(Nullpunkt) zwischen authentisch und nicht-authentisch steht zur Diskussion.
Zur Visualisierung wurde eine Kombination aus scharfen Kanten und linearen Verläufen gewählt, weshalb der entstehende Graph nicht als mathematisch beschreibbare Funktion zu werten ist. Es geht vielmehr um die unterschiedlich markierten Bereiche. Sie wurden in Anlehnung an amerikanische und deutsche Schulnotensysteme mit Zeichen bzw. Ziffern gekennzeichnet. Zusätzlich hat die Autorin den Versuch unternommen, passende Begrifflichkeiten zu finden (siehe Legende).
Der linke Bereich „echt“ ist sehr schmal, und orientiert sich an der bildforensischen Auffassung von Authentizität: Nahezu unbearbeitete Bilddateien, die eine direkte, digitale oder digitalisierte Fotografie enthalten, sind authentisch. Es folgt eine etwas weiter gefasste Kategorie „angepasst“, welche im journalistischen Sinne erlaubte Bearbeitungen enthält.
Der linear absinkende Bereich C könnte nochmals unterteilt werden (daher die Auslassung von D/4/grundsätzlich nicht falsch), beschreibt insgesamt jedoch die Kategorie von Bildern, die zwar sichtbar oder erahnbar ausschließlich durch Menschen „bearbeitet“, jedoch in ihrer Aussage nicht verfälscht wurden. Dieser Graubereich trägt dem Umstand Rechnung, dass die individuelle Einschätzung oft kontextbezogen erfolgt.
Im weiß gelassenen Bereich um den Nullpunkt sind Manipulationen zu verorten, da sie sowohl mit authentischen Bildern durch Rekontextualisierungen, als auch durch eine zu starke Bearbeitung erfolgen können.
Nach Auffassung der Autorin führen auch elementweise „Veränderungen mittels KI“ zu einer negativen Einschätzung hinsichtlich der Authentizität des Bildes. Durch den stetig wachsenden Einsatz von *machine learning* sogar in einfachen Bearbeitungsschritten wie dem Weichzeichnen des Hintergrunds[18], muss diese Einschätzung zukünftig gegebenenfalls revidiert werden, um allgemeine Zustimmung und praktische Umsetzung zu erfahren.
Ist ein Bild vollständig oder überwiegend „KI-generiert“, so sollte es allenfalls verifizierbar, jedoch keinesfalls authentifizierbar sein.

Um dem NutriScore näher zu kommen, wurde zusätzlich eine stark vereinfachte Skala erstellt, die in Abbildung 6.4 zu sehen ist. Sie enthält einen roten, mit M gekennzeichneten

[18] Wenn die Kamera-App mittels *machine learning* den Hintergrund ohne Zutun der Nutzer:in weichzeichnet, stellt das Bild noch eine authentische Szene dar, durch die KI-Bearbeitung würde es in der vorgestellten Skala jedoch in den Bereich nicht-authentischer Bilder fallen.

Abbildung 6.4: vereinfachte Skala in Anlehnung an den NutriScore; eigene Darstellung

Bereich für manipulierte, also nicht-authentische Bilder; KI-generierte Bilder erhalten die wertneutrale Farbe blau, während die anderen Farben an das Ampel-System angelehnt sind, um möglichst einfach verständlich zu sein. Bei weltweiter Implementation sollte die Leserichtung der betrachtenden Person berücksichtigt werden.

6.6 Resümee

Die Beantwortung der eingangs aufgestellten Forschungsfrage ist angesichts der Komplexität des Themas schwierig. Begrenzt auf einen journalistischen Kontext, unter Zuhilfenahme der geisteswissenschaftlichen und allgemeinen Interpretation des Konstruktes Authentizität muss die Frage zweigeteilt beantwortet werden.

Ja, Content Credentials können im Idealfall bei der Authentifizierung von Bildern hilfreich sein. Sind genügend Informationen über den Erstellungs- und Bearbeitungsprozess eines digitalen Bildes in den Metadaten gespeichert, kann leichter ermittelt werden, ob die Bildaussage durch Bearbeitung verfälscht wurde und damit eine Manipulation vorliegt. Auch Citizen Media kann leichter verifiziert werden, wenn Content Credentials vorhanden sind und von Journalist:innen als vertrauenswürdige Informationsquelle wahrgenommen werden. Vorausgesetzt, die Smartphone-und App-Entwicklung berücksichtigt die Spezifikation, könnten so durch die Initiative der CAI Citizen Media vermehrt Einzug in Rechtsprozesse und öffentliche Medien erhalten.

Und Nein, auch auf lange Sicht wird es angesichts der komplexen Spezifikation, ihrer Schwachstellen und Risiken und damit verbundener Möglichkeiten, manipulierte Bilder authentisch aussehen zu lassen, notwendig sein, eine zusätzliche bildforensische Analyse durchzuführen, um Bilder zu authentifizieren. Außerdem müssen Bildaussagen auch

verifiziert werden; hierbei erzeugen C2PA-Metadaten keinen weiteren Vorteil.

Viele Schwächen sind nicht nur auf fehlerhafte Implementationen zurückzuführen und die Risiken nicht nur in der Einführungsphase zu erwarten. Das führt dazu, dass Unternehmen bei der Umsetzung zögern und Plattformen die CCr nicht anzeigen. Außerdem müssen alle Menschen den Umgang mit Metadaten erlernen, denn ohne die Mitarbeit aller Beteiligten und der gesamten Bevölkerung wäre der Mehrwert der C2PA-Metadaten dem Aufwand nicht entsprechend. Selbst wenn die Implementation perfekt, und die Allgemeinheit unterrichtet wäre, bliebe die Verwendung des Konstrukts Authentizität durch die CAI fragwürdig. Sie misst digitale Fotografien und KI-generierte Bilder mit einerlei Maß, und zwar ausschließlich auf technischer Ebene. Dadurch rückt die bedeutende inhaltliche Ebene in den Hintergrund. Eine stärkere Differenzierung wie in Abschnitt 6.5 vorgeschlagen kann für mehr Klarheit sorgen. Dennoch bleibt das Vertrauensproblem bestehen, es verlagert sich lediglich von Journalist:innen und Medienschaffenden hin zu Software und Algorithmen.
Ob Content Credentials auch von Rezipient:innen als hilfreich wahrgenommen werden, hängt stark von der Implementierung der Verify-Seiten und dem Willen der Individuen ab, sich mit zusätzlichen Informationen auseinanderzusetzen.

Vor diesem Hintergrund ist es erstaunlich, dass im November 2024 das NIST berichtete, dass das ISO Technical Committee[19] 171 mit dem Entwurf ISO/CD 22144 daran arbeitet, die C2PA-Spezifikation V. 2.1 in einen Standard zu überführen (vgl. auch NSA u.a. 2025). Externe Akteure kritisieren, dass die Spezifikation sowie ihre Open-Source-Implementationen nicht ausreichend von unabhängigen Instanzen überprüft wurden (vgl. Krawetz 2024a; NIST 2024). Eine Standardisierung könnte jedoch der Verbreitung und systemweiten Implementation einen Schub geben und so die Wirksamkeit erhöhen.

Auf den Punkt gebracht: Die Vision der CAI ist wegweisend, die technische Umsetzung jedoch noch fehlerbehaftet und ihre Wirksamkeit stark abhängig von gesellschaftlichen und individuellen Faktoren, die außerhalb des Einflussbereichs der beteiligten Unternehmen liegen. Authentifizierung und Verifizierung von Bildern bleibt auch mit CCr ein von Menschen geleiteter Prozess, der durch C2PA-Metadaten und Algorithmen zwar verbessert oder vereinfacht, aber keinesfalls ersetzt werden kann.

[19] Vorsitzender des Komitees ist Leonard Rosenthol (Adobe).

6.7 Reflexion

Durch die Kombination aus Recherchearbeit, praktischen Versuchen und der Einbeziehung verschiedener Blickwinkel konnte die Komplexität des Themenfeldes dargestellt werden. Eine getrennte Betrachtung hätte nahegelegt, dass die Bildforensik keine zusätzlichen Metadaten braucht, und die Nachrichtenredaktionen verstärkt auf OSINT Technologien setzen und ihre Arbeitsweisen anpassen können, um aktuellen Herausforderungen zu begegnen. Erst durch die gemeinsame Betrachtung aller Aspekte und Teilbereiche zeigte sich das Potenzial der C2PA-Spezifikation.
Es wurde bewusst entschieden, englische Fachbegriffe zu übersetzen und den Originalbegriff beizubehalten, um sowohl Verständlichkeit als auch Nachvollziehbarkeit zu gewährleisten.
Auf Grund der Aktualität der C2PA-Spezifikation und der begrenzten Versuchszeit haben die Versuche lediglich eine erste Einschätzung und Beurteilung ermöglicht. Es ist nicht klar geworden, warum auf Social-Media-Plattformen trotz Ankündigung keine CCr zu finden waren.
Interviews mit bereits involvierten Personen im WDR und der DW sowie großen Verlagen hätte den zeitlichen Rahmen gesprengt, wären aber gerade bei der Einschätzung im Hinblick auf das Potenzial im journalistischen Kontext hilfreich gewesen. Auch ein dezidierter Blick auf alternative Arbeitsweisen zur Differenzierung von synthetischen und nicht-synthetischen Bilder innerhalb der Branche und der Umgang mit Rekontextualisierungen hätte den Rahmen gesprengt.
Um das Vertrauensmodell der C2PA besser verstehen und einordnen zu können, hätte es zusätzlich einer intensiven Einarbeitung in Themen wie „web of trust“ und PKI bedurft.

Die Diskussionen zu den Themen Authentizität, Falschinformationen, und möglicher technischer Lösungen zeigen, dass der Ansatz der C2PA *eine* mögliche Lösung ist. Alternative, weniger öffentlich diskutierte Methoden und auch technische Optionen lassen vermuten, dass es weitere Entwicklungen geben wird, die mit den CCr konkurrieren, oder sie ergänzen werden. Zugleich darf die soziologische, ökonomische und politische Perspektive auf das Thema nicht vernachlässigt werden. Es bleibt zukünftigen Arbeiten überlassen, die Wirkung von Herkunftsnachweisen oder ähnlichen Zusatzinformationen mit Hilfe von Studien zu untersuchen und ihre Anwendung gegebenenfalls zu verbessern.

Anhang

A Anlagen zu den Versuchen

Für die Versuche wurde folgende Software verwendet:

- Exiftool Version 13.24
- c2patool Version 0.14.0
- Windows-Terminal Version: 1.22.10731.0
- Adobe Photoshop Version 26.5
- GIMP Version 3.0.2
- Libre Office Calc Version 24.8.2.1
- Obsidian Version 1.7.4 (Protokollnotizen erstellen)

Die Auswahl der verwendeten Software erfolgte pragmatisch. Für den Versuch wurde eine personengebundene, Nicht-Education-Lizenz von Adobe Photoshop erworben[1]. Das https://exiftool.org/ von Phil Harvey ist open source, plattformübergreifend, über die Kommandozeile nutzbar, frei verfügbar und kann lokal betrieben werden. Außerdem zeigt es C2PA-Metadaten an, ist gut dokumentiert und wird laufend weiterentwickelt. Der Service Hintfo[2] bedient sich des Exiftools, sodass Leser:innen auch ohne Kommandozeile die Vesuchsbilder unkompliziert untersuchen können.
Das https://github.com/contentauth/c2patool wurde bis Dezember 2024 von der CAI ge-

[1] Education-Lizenzen ermöglichten zum Zeitpunkt der Bearbeitung (April 2025) noch nicht die Content Credential Funktionalität, wenngleich bereits bestehende CCr aus KI-generierten Bildern in Photoshop mit einer Education-Lizenz bearbeitet werden konnten, ohne dass die CCr zerstört wurden, wie Vorversuche ergaben.
[2] Der Service wird von Neal Krawetz' Hacker Factor betrieben.

pflegt und danach in die Sprache Rust umgewandelt. Von der neuesten Version c2pa-rs gibt es zum Zeitpunkt der Bearbeitung keine ausführbare Version, sodass mit der Version aus Dezember 2024 gearbeitet wird.
Das GIMP von GNU ist ein sehr verbreitetes open source Bildbearbeitungsprogramm, hat die C2PA-Spezifikation jedoch zum Zeitpunkt der Bearbeitung noch nicht implementiert. Dadurch eignet es sich zur Analyse, wie nicht C2PA-kompatible Software mit C2PA-Metadaten umgeht.
Zu Versuchbeginn hat Microsoft seine Verify-Seite `https://contentintegrity.microsoft.com/check` an einen Login mit Microsoft-Konto gebunden. Mit einem einfachen Konto ist der Zugang zur Seite nicht (mehr) möglich. Deswegen wurde sich ausschließlich auf die Verify-Seite der CAI bezogen: `https://contentcredentials.org/verify`ProofMode arbeitet nach einem eigenen Prinzip, welches ein eigenes Verifkationswerkzeug erforderlich macht: `https://check.proofmode.org/`.
Die Auswahl der Hardware erfolgte ebenfalls pragmatisch; zugleich jedoch realitätsnahe. Es bestand kein Zugriff auf eine C2PA-fähige Kamera. Wichtig für die Beantwortung der Hypothesen ist, dass unterschiedliche Metadatensätze zur Verfügung stehen. Es wurden alle Metadaten aufzuzeichnen, um eine größere Flexibilität in der Analyse und Auswertung zu erlangen.

B Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit mit dem Titel

Authentifizierung digitaler Bilder durch Metadaten und Bildforensik vor dem Hintergrund KI-generierter Bilder und Citizen Media im Kontext redaktioneller Arbeit

selbstständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Alle Passagen, die ich wörtlich aus der Literatur oder aus anderen Quellen wie z.B. Internetseiten übernommen habe, habe ich deutlich als Zitat mit Angabe der Quelle kenntlich gemacht.

Lübeck, 2. Juni 2025

Literatur

Apple - September Event (7. Sep. 2016).
URL: https://www.youtube.com/watch?v=NS0txu_Kzl8 (besucht am 23. 04. 2025).

Archive, Syrian (13. Sep. 2016).
Dataset of Verified Videos About Chemical Weapons Attacks in Syria. bellingcat.
URL: https://www.bellingcat.com/news/mena/2016/09/13/dataset-verified-videos-chemical-weapons-attacks-syria/ (besucht am 14. 04. 2025).

Associated Press (2025). *Telling the Story*. Associated Press. URL:
https://www.ap.org/about/news-values-and-principles/telling-the-story/
(besucht am 01. 05. 2025).

AzSoS Certified Media (2025). URL: https://acm.azsos.gov/ (besucht am 20. 03. 2025).

Balan, Kar, Shruti Agarwal, Simon Jenni, Andy Parsons, Andrew Gilbert und John Collomosse (Juni 2023). „EKILA: Synthetic Media Provenance and Attribution for Generative Art“.
In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, S. 913–922. DOI: 10.1109/CVPRW59228.2023.00098.

Bharati, Aparna, Daniel Moreira, Joel Brogan, Patricia Hale, Kevin W. Bowyer, Patrick J. Flynn, Anderson Rocha und Walter J. Scheirer (6. März 2019).
Beyond Pixels: Image Provenance Analysis Leveraging Metadata.
DOI: 10.48550/arXiv.1807.03376. arXiv: 1807.03376 [cs].
URL: http://arxiv.org/abs/1807.03376 (besucht am 11. 04. 2025).
Vorveröffentlichung.

Bhowmik D, Caldwell S, Delgado J, Ebrahimi T, Fotos N, Gu X, Hu Z, Kang X, Pereira F, Rosenthol L und Temmermans F (25. Okt. 2024).
„An International Standard For Assessing Trustworthiness In Media“.
In: *IEEE International Conference on Image Processing (ICIP)*. IEEE Explore, S. 3799–3805. DOI: 10.57711/am2a-ad03.

BSI Glossar (2025).
Bundesamt für Sicherheit in der Informationstechnik - Glossar - Hashfunktion.

Bundesamt für Sicherheit in der Informationstechnik.
URL: https://www.bsi.bund.de/DE/Themen/Oeffentliche-Verwaltung/Moderner-Staat/ElektronischeSignatur/Glossar/esigglossar.html?nn=450092#Glossar_H (besucht am 19. 01. 2025).

Bundesdruckerei GmbH (2024). *Bundesdruckerei tritt C2PA bei*.
URL: https://www.bundesdruckerei.de/de/newsroom/news/bundesdruckerei-tritt-c2pa-bei (besucht am 24. 01. 2025).

– (2025). *Content Credentials - Schutz vor Bildmanipulationen im KI-Zeitalter*. URL: https://www.bundesdruckerei.de/de/innovation-hub/content-credentials-schutz-vor-bildmanipulationen-im-ki-zeitalter (besucht am 19. 01. 2025).

Bundeskriminalamt, Hrsg. (Nov. 2024).
Auf der Spur mit KI – wie KI die polizeiliche Welt revolutioniert.
URL: https://www.bka.de/SharedDocs/Downloads/DE/Publikationen/Publikationsreihen/CodLiteraturreihe/8_33_Auf_der_Spur_mit_KI.pdf?__blob=publicationFile&v=2 (besucht am 28. 03. 2025).

Butora, Jan und Patrick Bas (24. Juni 2024).
„The Adobe Hidden Feature and Its Impact on Sensor Attribution“.
In: *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*. IH&MMSec '24. New York, NY, USA: Association for Computing Machinery, S. 143–148. ISBN: 979-8-4007-0637-0. DOI: 10.1145/3658664.3659650.

C2PA (1. Juni 2022).
C2PA Harms, Misuse, and Abuse Assessment; Phase III - Existing and Potential Mitigations.
URL: https://c2pa.org/specifications/specifications/2.0/security/_attachments/Due_Diligence_Actions.pdf (besucht am 11. 05. 2025).

– (20. Sep. 2024). *Content Credentials: C2PA Technical Specification*.
URL: https://c2pa.org/specifications/specifications/2.1/specs/C2PA_Specification.html (besucht am 14. 01. 2025).

– (2025). *C2PA Harms Modelling*. URL: https://c2pa.org/specifications/specifications/2.0/security/Harms_Modelling.html (besucht am 11. 05. 2025).

C2PA Implementation Strategies (29. Jan. 2025). Genf.
URL: https://tech.ebu.ch/publications/presentations/2025/pts2025/c2pa-implementation-strategies (besucht am 27. 03. 2025).

Castellanos, Jacobo und Sarah Gulliford (Kearns) (19. Sep. 2022). „Layers of Trust“.
In: *Commonplace*. DOI: 10.21428/6ffd8432.3b76421c.

Clegg, Nick (6. Feb. 2024).
Labeling AI-Generated Images on Facebook, Instagram and Threads. Meta Newsreoom.
URL: https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/ (besucht am 24. 03. 2025).

Collomosse, John und Andy Parsons (Mai 2024a). „To Authenticity, and Beyond! Building Safe and Fair Generative AI Upon the Three Pillars of Provenance“.
In: *IEEE Computer Graphics and Applications* 44.3, S. 82–90. ISSN: 0272-1716, 1558-1756. DOI: 10.1109/MCG.2024.3380168.

– (Mai 2024b). „To Authenticity, and Beyond! Building Safe and Fair Generative AI Upon the Three Pillars of Provenance“. In: *IEEE Computer Graphics and Applications* 44.3, S. 82–90. ISSN: 1558-1756. DOI: 10.1109/MCG.2024.3380168.
URL: https://ieeexplore.ieee.org/document/10568485/?arnumber=10568485 (besucht am 20. 03. 2025).

Content Authenticity Initiative (2025a). *Getting Started with Content Credentials | Open-source Tools for Content Authenticity and Provenance*. Getting started with Content Credentials.
URL: https://opensource.contentauthenticity.org/docs/getting-started/ (besucht am 09. 01. 2025).

Content Credentials im Überblick (12. Feb. 2025). Content Credentials im Überblick. URL: https://helpx.adobe.com/content/help/de/de/firefly/get-set-up/learn-the-basics/content-credentials-overview.html (besucht am 24. 03. 2025).

Corrigan, Patrick (15. Mai 2024). *LinkedIn Adopts C2PA Standard*.
LinkedIn Adopts C2PA Standard. URL: https://www.linkedin.com/pulse/linkedin-adopts-c2pa-standard-patrick-corrigan-kwldf (besucht am 24. 03. 2025).

CYBERGEM [@UltraTerm] (12. Aug. 2024).
I Have Been BANNED from @Adobe's C2PA Discord Server!
Screenshots des Posts sind auf dem USB-Stick im Ordner Quellen zu finden. X. URL: https://x.com/UltraTerm/status/1823077457395953866 (besucht am 21. 05. 2025).

Deutscher Presserat (19. März 2025). *Pressekodex - Ethische Standards für den Journalismus*. Bearb. von Deutscher Presserat.
URL: https://www.presserat.de/pressekodex.html (besucht am 24. 04. 2025).

Dillon, Conor (6. Jan. 2025). *Fake MH17 Satellite Pics*. dw.com.
URL: https://www.dw.com/en/forensic-report-russia-faked-mh17-satellite-photos/a-18490259 (besucht am 14. 04. 2025).

Ellis, Laura (22. Apr. 2022). *Project Origin: Securing Trust in Media*.
Project Origin: Securing Trust in Media - Beyond Fake News.

URL: https://www.bbc.com/beyondfakenews/trusted-news-initiative/bbc.com/beyondfakenews/trusted-news-initiative/project-origin-securing-trust-in-media/ (besucht am 15. 03. 2025).

England, Paul, Henrique S. Malvar, Eric Horvitz, Jack W. Stokes, Cédric Fournet, Rebecca Burke-Aguero, Amaury Chamayou, Sylvan Clebsch, Manuel Costa, John Deutscher, Shabnam Erfani, Matt Gaylor, Andrew Jenks, Kevin Kane, Elissa Redmiles, Alex Shamis, Isha Sharma, Sam Wenker und Anika Zaman (20. Juni 2020). *AMP: Authentication of Media via Provenance*. DOI: 10.48550/arXiv.2001.07886. arXiv: 2001.07886 [cs.MM]. Vorveröffentlichung.

Feng, K. J. Kevin, Nick Ritchie, Pia Blumenthal, Andy Parsons und Amy X. Zhang (28. Sep. 2023). „Examining the Impact of Provenance-Enabled Media on Trust and Accuracy Perceptions“. In: *Proceedings of the ACM on Human-Computer Interaction* 7 (CSCW2), S. 1–42. ISSN: 2573-0142. DOI: 10.1145/3610061.

Gangwar, D P und Anju Pathania (1. Dez. 2018). „Authentication of Digital Image Using Exif Metadata and Decoding Properties“. In: *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 3.8, S. 335–341. DOI: 10.32628/CSEIT183815. URL: https://www.researchgate.net/publication/329880328_Authentication_of_Digital_Image_using_Exif_Metadata_and_Decoding_Properties (besucht am 11. 04. 2025).

Gerling, Winfried (24. Feb. 2022). „Bildforensik im Journalismus- Kontexte und Methoden“. In: *Fotojournalismus im Umbruch: Hybrid, multimedial, prekär*. 1. Auflage. Köln: Herbert von Halem Verlag, S. 296–317. ISBN: 978-3-86962-559-1. DOI: 10.1453/2022_9783869625591.

– (Juli 2024). „Das Bild als Wahrscheinlichkeit“. In: *Bild | Kanäle. Zur Theorie und Ästhetik vernetzter Medienkultur*. Königshausen & Neumann, S. 39–70. ISBN: 978-3-8260-7373-1.

Gerling, Winfried, Susanne Holschbach und Petra Löffler (2018). *Bilder verteilen: fotografische Praktiken in der digitalen Kultur*. Bd. Digitale Gesellschaft. Digitale Gesellschaft 18. Bielefeld: transcript. 287 S. ISBN: 978-3-8376-4070-0.

Gerth, Sebastian (3. Juni 2018). „Auf Der Suche Nach Visueller Wahrheit: Authentizitätszuschreibung Und Das Potenzial Der Wirklichkeitsabbildung Durch Pressefotografien Im Zeitalter Digitaler Medien“. In.

Gray, Jeremy (3. Feb. 2025).
Cloudflare Joins CAI and Enables Content Credentials for 20% of the Internet. PetaPixel.
URL: https://petapixel.com/2025/02/03/cloudflare-joins-cai-and-enables-content-credentials-for-20-of-the-internet/ (besucht am 27. 03. 2025).

Grittmann, Elke (2003). „Die Konstruktion von Authentizität. Was ist echt an den Pressefotos im Informationsjournalismus?“ In: *Authentizität und Inszenierung von Bilderwelten*.
Köln: Herbert von Halem Verlag, S. 123–149. ISBN: 978-3-931606-49-7.

Inc., Adobe (2025). *Inhaltsurhebernachweise (Beta) in Photoshop.*
URL: https://www.adobe.com/de/learn/photoshop/web/secure-image-data-enable-content-credentials-photoshop (besucht am 16. 03. 2025).

Information Technology Industry Council (1. Jan. 2024). *Authenticating AI-Generated Content, Exploring Risks, Techniques & Policy Recommendations*. ITI. URL: https://www.itic.org/policy/ITI_AIContentAuthorizationPolicy_122123.pdf (besucht am 15. 01. 2025).

Initiative, Content Authenticity (13. Juni 2023). *CONTRIBUTING.md*.
URL: https://github.com/contentauth/c2pa-rs/blob/main/CONTRIBUTING.md (besucht am 07. 02. 2025).

– (9. Apr. 2025b). *Cargo.toml*. C2PA.
URL: https://github.com/contentauth/c2pa-rs/blob/main/sdk/Cargo.toml (besucht am 14. 04. 2025).

IPTC (8. Mai 2024). *Panel 1: AI and Image Authenticity*.
URL: https://www.youtube.com/watch?v=4q7iBkRLCMQ (besucht am 12. 04. 2025).

Kirchner, Matthias (2022). „Sensor Fingerprints: Camera Identification and Beyond“.
In: *Multimedia Forensics*.
Hrsg. von Husrev Taha Sencar, Luisa Verdoliva und Nasir Memon.
Singapore: Springer Singapore, S. 65–88. ISBN: 978-981-16-7621-5.
DOI: 10.1007/978-981-16-7621-5_4.

Klussmann, Niels (2000). *Lexikon der Kommunikations- und Informationstechnik: Telekommunikation, Datenkommunikation, Multimedia, Internet*.
2., erw. und aktualisierte Aufl. Heidelberg: Hüthig. 871 S. ISBN: 978-3-7785-3913-2.

Krämer, Benjamin und Katharina Lobinger (2018).
„’So und nicht anders ist es gewesen!’: Visuelle Authentizitäten und die Rolle kontextspezifischer Authentizitätsmarker in der visuellen Kommunikation“.
In: *Handbuch Visuelle Kommunikationsforschung*. Hrsg. von Katharina Lobinger.

Wiesbaden: Springer Fachmedien Wiesbaden, S. 1–21. ISBN: 978-3-658-06738-0. DOI: 10.1007/978-3-658-06738-0_6-1.

Krawetz, Neal (31. Dez. 2024a). *Another Year Down*. The Hacker Factor Blog. URL: https://www.hackerfactor.com/blog/index.php?/archives/1053-Another-Year-Down.html (besucht am 14. 04. 2025).

– (15. Okt. 2024b). *C2PA and Authenticated Disinformation*. URL: https://hackerfactor.com/blog/index.php?/archives/1046-C2PA-and-Authenticated-Disinformation.html (besucht am 20. 03. 2025).

– (9. Aug. 2024c). *C2PA and the CBC*. The Hacker Factor Blog. URL: https://hackerfactor.com/blog/index.php?/archives/1040-C2PA-and-the-CBC.html (besucht am 03. 02. 2025).

– (9. Mai 2024d). *C2PA from the Attacker's Perspective*. The Hacker Factor Blog. URL: https://www.hackerfactor.com/blog/index.php?/archives/1031-C2PA-from-the-Attackers-Perspective.html (besucht am 27. 03. 2025).

– (18. Apr. 2025). *C2PA and Authentication Updates*. The Hacker Factor Blog. URL: https://hackerfactor.com/blog/index.php?/archives/1065-C2PA-and-Authentication-Updates.html (besucht am 01. 05. 2025).

Kübler, Hans-Dieter (2022). *Bildjournalismus und Pressefotografie: Geschichte, mediale Formate, Analysen: eine Einführung*. Wiesbaden: Springer VS. ISBN: 978-3-658-35291-2. DOI: 10.1007/978-3-658-35292-9.

Leifert, Stefan (2007). *Bildethik: Theorie und Moral im Bildjournalismus der Massenmedien*. München: Wilhelm Fink. ISBN: 978-3-7705-4416-5.

Loosen, Wiebke und Paul Solbach (2020). „Daten und Algorithmen". In: *Fake News, Framing, Fact-Checking - Nachrichten im digitalen Zeitalter*. Bd. 30. Digitale Gesellschaft. Bielefeld: transcript Verlag, S. 554. ISBN: 978-3-8376-5025-9.

Lorch, Benedikt (27. Feb. 2023). „Reliable Machine Learning Methods in Image Forensics". Nürnberg: Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). URL: https://open.fau.de/items/cd0ff7f3-196a-49fc-a73c-283e0d5b1e6a (besucht am 18. 03. 2025).

Masood, Adnan Masood (29. März 2025). *Toward Reliable Provenance in AI-Generated Content: Text, Images, and Code*. Medium. URL: https://medium.com/@adnanmasood/toward-reliable-provenance-in-ai-generated-content-text-images-and-code-9ebe8c57ceae (besucht am 30. 04. 2025).

National Institute of Standards and Technology (18. Nov. 2024). *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency*. NIST AI NIST AI 100-4. Gaithersburg, MD: National Institute of Standards and Technology, S. 81. DOI: 10.6028/NIST.AI.100-4. (Besucht am 20. 03. 2025).

National Security Agency, USA,
Australien Signals Directorate's Australien Cyber Security Centre,
Canadian Centre for Cyber Security und United Kindowm Nationa Cyber Security Centre (Jan. 2025).
Content Credentials: Strengthening Multimedia Integrity in the Generative AI Era.
URL: https://media.defense.gov/2025/Jan/29/2003634788/-1/-1/0/CSI-CONTENT-CREDENTIALS.PDF (besucht am 11. 04. 2025).

NewsCodes Scheme (2024). *NewsCodes Scheme for Digital Source Type*. URL: https://cv.iptc.org/newscodes/digitalsourcetype/ (besucht am 26. 03. 2025).

Nightingale, Sophie J. und Hany Farid (22. Feb. 2022).
„AI-synthesized Faces Are Indistinguishable from Real Faces and More Trustworthy".
In: *Proceedings of the National Academy of Sciences* 119.8, e2120481119.
ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2120481119.

Okeowo, Adebayo (14. Okt. 2022). „Advancing accountability for human rights violations through citizen media - an African perspective". Diss. University of Pretoria. 218 S.
URL: https://repository.up.ac.za/handle/2263/88544 (besucht am 04. 02. 2025).

OpenAI (2025). *C2PA in DALL·E 3 | OpenAI Help Center*. Help Center Artikel.
URL: https://help.openai.com/en/articles/8912793-c2pa-in-dall-e-3 (besucht am 15. 01. 2025).

Orlosky, Mick (6. Mai 2024). *Press Release: Camera Bits Introduces Its Solution for Protecting Provenance of C2PA Signed Photos in Effort to Help Combat Fake Imagery*.
Camera Bits, Inc.
URL: https://home.camerabits.com/2024/05/06/press-release-camera-bits-introduces-its-solution-for-protecting-provenance-of-c2pa-signed-photos-in-effort-to-help-combat-fake-imagery/ (besucht am 19. 01. 2025).

Osing, Tim (2022). *Digitaler Journalismus in der Praxis: Grundlagen von Onlinerecherche, Storytelling und Datenjournalismus*. Wiesbaden: Springer VS. ISBN: 978-3-658-39105-8.
DOI: 10.1007/978-3-658-39105-8.

Parsons, Andy (10. Okt. 2023). *Adobe MAX 2023: Einführung des neuen Content Credential „Icon of Transparency" | Adobe Blog*. Adobe Inc.
URL: https://blog.adobe.com/de/publish/2023/10/10/adobe-max-2023-

einfuehrung-des-neuen-content-credential-icon-of-transparency (besucht am 24. 03. 2025).

Parsons, Andy (31. Jan. 2024a). *Die Gunst der Stunde nutzen: die Akzeptanz von Content Credentials im Jahr 2024 vorantreiben*. Adobe Experince Cloud. URL: https://business.adobe.com/de/blog/neues/die-akzeptanz-von-content-credentials-im-jahr-2024-vorantreiben (besucht am 03. 02. 2025).

– (8. Apr. 2024b). *Durable Content Credentials*. Content Authenticity Initiative. URL: https://contentauthenticity.org/blog/durable-content-credentials (besucht am 01. 05. 2025).

Peters, Martha Rebekka (14. Juli 2023). „Die Verifikation audiovisuellen Materials im digitalen Zeitalter. Technische Möglichkeiten und journalistische Praxis“. Magisterarb. Hochschule Bonn-Rhein-Sieg. 132 S. URL: https://pub.h-brs.de/frontdoor/index/index/docId/7502 (besucht am 24. 01. 2025).

PhotoShelter (17. Sep. 2024). *What Are Content Credentials?* Q: What are Content Credentials. URL: https://go.photoshelter.com/ask-photoshelter/what-are-content-credentials/ (besucht am 12. 05. 2025).

Piva, Alessandro und Massimo Iuliani (2022). „Integrity Verification Through File Container Analysis“. In: *Multimedia Forensics*. Hrsg. von Husrev Taha Sencar, Luisa Verdoliva und Nasir Memon. Singapore: Springer Singapore, S. 363–387. DOI: 10.1007/978-981-16-7621-5_14.

Project Origin (2025). *Project Origin - Protecting Trusted Media*. Project Origin. URL: https://www.originproject.info (besucht am 13. 01. 2025).

Ratering, Jörn (31. Aug. 2022). „Zwischen Geotags und Gesichtserkennung: KI in der Verifikation“. In: *Communicatio Socialis (ComSoc)* 55.3, S. 360–366. ISSN: 0010-3497, 2198-3852. DOI: 10.5771/0010-3497-2022-3-360.

Rathi, Kaushal, Sathyanarayana Sampath Kumar und A. N. Mandanna (1. Feb. 2024). *Insights into Coalition for Content Provenance and Authenticity (C2PA)*. Infosys Tech Compass. URL: https://www.infosys.com/iki/techcompass/content-provenance-authenticity.html (besucht am 17. 03. 2025).

Riess, Christian (2022). „Physical Integrity“. In: *Multimedia Forensics*. Hrsg. von Husrev Taha Sencar, Luisa Verdoliva und Nasir Memon. Singapore: Springer, S. 207–234. ISBN: 978-981-16-7621-5. DOI: 10.1007/978-981-16-7621-5_9.

Rossig, Julian J. (2014). *Fotojournalismus*. 3., völlig überarb. Aufl. Praktischer Journalismus. Konstanz München: UVK-Verlagsgesellschaft. 266 S. ISBN: 978-3-86764-482-2.

Rothöhler, Simon (2024). „Ikonische Indifferenz. Zur Bildlosigkeit der Bildforensik“. In: *Bilder unter Verdacht: Praktiken der Bildforensik*. Hrsg. von Roland Meyer. Bildwelten des Wissens Band 19. Berlin/Boston: De Gruyter, S. 34–46. ISBN: 978-3-11-108569-2. DOI: 10.1515/9783111085692001.

Russezki, Jan und Maike Schultz (20. März 2019). „AfD verbreitet manipuliertes Foto von Klimademo“. In: *RND*. URL: https://www.rnd.de/medien/afd-verbreitet-manipuliertes-foto-von-klimademo-7OHHYT2V7A7YL4OGQSQHPXKNY4.html (besucht am 23. 04. 2025).

Ryan-Mosley, Tate (6. Nov. 2023). *The inside Scoop on Watermarking and Content Authentication*. MIT Technology Review. URL: https://www.technologyreview.com/2023/11/06/1082996/the-inside-scoop-on-watermarking-and-content-authentication/ (besucht am 12. 04. 2025).

Saberi, Mehrdad, Vinu Sankar Sadasivan, Arman Zarei, Hessam Mahdavifar und Soheil Feizi (20. Juni 2024). *DREW: Towards Robust Data Provenance by Leveraging Error-Controlled Watermarking*. DOI: 10.48550/arXiv.2406.02836. Vorveröffentlichung.

Salisbury, Meredith und Jefferson Pooley (20. Jan. 2017). „The #nofilter Self: The Contest for Authenticity among Social Networking Sites, 2002–2016“. In: *Social Sciences* 6, S. 10. DOI: 10.3390/socsci6010010.

Scientific Working Group on Digital Evidence, Hrsg. (3. März 2025). *Best Practices for Image Authentication*. URL: https://www.swgde.org/18-i-001/ (besucht am 18. 03. 2025).

Simmons, John C und Joseph M Winograd (20. Mai 2024). *Interoperable Provenance Authentication of Broadcast Media Using Open Standards-based Metadata, Watermarking and Cryptography*. arXiv: 2405.12336 [cs.CR]. URL: https://arxiv.org/abs/2405.12336 (besucht am 14. 01. 2025). Vorveröffentlichung.

Steadman, Ian (16. Mai 2013). „’Fake’ World Press Photo Isn’t Fake, Is Lesson in Need for Forensic Restraint“. In: *Wired*. URL: https://www.wired.com/story/photo-faking-controversy/ (besucht am 23. 04. 2025).

Stephan, Julia (31. Juli 2024). *Content Verification: A Project for Photo Authenticity in Journalism*. Fotoware.

URL: https://www.fotoware.com/blog/content-verification-photo-authenticity-journalism (besucht am 12. 05. 2025).

Stern, Jenny (3. Juni 2020). „Fact-Checking und Verifikation“. In: *Fake News, Framing, Fact-Checking: Nachrichten im digitalen Zeitalter: Ein Handbuch*. Hrsg. von Köhler, Tanja. 1. Auflage. Digitale Gesellschaft. Bielefeld: transcript Verlag, S. 119–150. ISBN: 978-3-8394-5025-3. DOI: 10.1515/9783839450253.

Technische Richtlinie TR-03183: Cyber-Resilienz-Anforderungen an Hersteller und Produkte (12. Juli 2023). Techn. Ber. Bundesamt für Sicherheit in der Informationstechnik. URL: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Publikationen/TechnischeRichtlinien/TR03183/BSI-TR-03183-2.pdf?__blob=publicationFile&v=3 (besucht am 14. 04. 2025).

Temmermans, F., S. Caldwell, D. Bhowmik und T. Ebrahimi (1. Okt. 2024). „JPEG Trust: An International Standard Facilitating the Assessment of Trustworthiness of Digital Media Assets“. In: Applications of Digital Image Processing XLVII. SPIE. DOI: 10.57711/KXK8-DK55.

The News Provenance Project (2025). The News Provenance Project. URL: https://www.newsprovenanceproject.com (besucht am 13. 01. 2025).

TikTok (9. Mai 2024). *Partnering with Our Industry to Advance AI Transparency and Literacy*. TikTok. URL: https://newsroom.tiktok.com (besucht am 25. 03. 2025).

truepic (23. Apr. 2025). *What happens if real is actually fake?* URL: https://www.truepic.com/learning/transparency-in-ai (besucht am 24. 04. 2025).

Walsh, Joan (15. März 2006). „The Abu Ghraib Files“. In: *Der Spiegel*. ISSN: 2195-1349. URL: https://www.spiegel.de/international/spiegel-surfs-the-web-the-abu-ghraib-files-a-406055.html (besucht am 14. 04. 2025).

Welchering, Peter (2020). *Journalistische Praxis: Verifikation und Fact Checking*. Essentials Ser. Wiesbaden: Springer Fachmedien Wiesbaden GmbH. 1 S. ISBN: 978-3-658-30976-3. DOI: 10.1007/978-3-658-30977-0.

Wen, Yuxin, John Kirchenbauer, Jonas Geiping und Tom Goldstein (4. Juli 2023). *Tree-Ring Watermarks: Fingerprints for Diffusion Images That Are Invisible and Robust*. DOI: 10.48550/arXiv.2305.20030. arXiv: 2305.20030 [cs]. Vorveröffentlichung.

Wolters, Sven und Claudia Bracholdt (7. Sep. 2018). *Das Chemnitz-Video im Faktencheck*. URL: https://www.zeit.de/video/2018-09/5832082264001/chemnitz-das-chemnitz-video-im-faktencheck (besucht am 26. 03. 2025).

Xiang, Ziyue, Janos Horvath, Sriram Baireddy, Paolo Bestagini, Stefano Tubaro und Edward J. Delp (Juni 2021). „Forensic Analysis of Video Files Using Metadata“. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Nashville, TN, USA: IEEE, S. 1042–1051. DOI: 10.1109/CVPRW53098.2021.00115.

YouTube, Google (2025). *Transparenz auf YouTube stärken: Offenlegung „Mit einer Kamera aufgenommen“*. YouTube. URL: https://support.google.com/youtube/answer/15446725?hl=de (besucht am 27. 03. 2025).

Zeloof, Adam (30. Nov. 2023). *Falsified Photos: Fooling Adobe's Cryptographically-Signed Metadata*. Hackaday. URL: https://hackaday.com/2023/11/30/falsified-photos-fooling-adobes-cryptographically-signed-metadata/ (besucht am 10. 04. 2025).

Zettelmeister, Joshua (9. Mai 2024). „Fotojournalistische Authentizität in Zeiten fotorealistischer Bildgenerierung“. Bachelorarbeit. Köln: TH Köln. 82 S. URL: https://publiscologne.th-koeln.de/frontdoor/deliver/index/docId/2508/file/BA_Zettelmeier_Joshua.pdf (besucht am 28. 01. 2025).

Zhao, Xuandong, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang und Lei Li (31. Okt. 2024). *Invisible Image Watermarks Are Provably Removable Using Generative AI*. DOI: 10.48550/arXiv.2306.01953. arXiv: 2306.01953 [cs]. Vorveröffentlichung.