

AI Watermarking Won't Curb Disinformation

Dokumentation von Quelltext und Übersetzung

Autor	Jacob Hoffman-Andrews
Originaltitel	AI Watermarking Won't Curb Disinformation
Übersetzungstitel	KI-Wasserzeichen sind kein wirksamer Schutz vor Desinformation
Quelle	EFF Deeplinks Blog (https://www.eff.org/deeplinks/2024/01/ai-watermarking-wont-curb-disinformation)
Lizenzangabe	CC BY 3.0 US

Zusammenfassung

Der Beitrag erläutert, warum Wasserzeichen in KI-generierten Bildern und Texten kaum geeignet sind, Desinformation wirksam einzudämmen. Er beschreibt technische Ansätze wie unsichtbare Bildmuster, Metadaten und stilistische Textmarkierungen, zeigt aber ihre Anfälligkeit gegenüber gezielter Entfernung, Umformulierung und alternativen KI-Systemen. Als Gegenansatz wird Content Authenticity erwähnt, die echte Aufnahmen kennzeichnen soll, jedoch ebenfalls keine umfassende Lösung bietet.

Summary

The article explains why watermarking AI-generated images and text is unlikely to effectively curb disinformation. It outlines technical approaches such as invisible image patterns, metadata, and subtle textual markers, but emphasizes their vulnerability to deliberate removal, rewording, and alternative AI systems. It also discusses content authenticity as a way to identify genuine camera images, while noting that this approach is not comprehensive either.

Übersetzungsherausforderungen

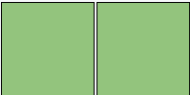
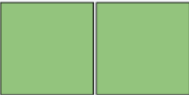
- Fachbegriffe wie „watermarking“, „content authenticity“, „metadata“ und „robustness“ mussten verständlich und zugleich nah am technischen Sinn wiedergegeben werden.
- Der Text verbindet technische Erklärung mit argumentativer Zuspitzung; die Übersetzung musste diesen sachlich-kritischen Ton ohne werbliche Glättung halten.
- Einige Beispiele und Referenzen, etwa The Federalist Papers oder Dienste wie GPTZero und TurnItIn, erfordern Kontexttreue, ohne den Text erklärend zu erweitern.

Translation Challenges

- Technical terms such as “watermarking,” “content authenticity,” “metadata,” and “robustness” had to remain understandable while preserving their technical meaning.
- The text combines technical explanation with pointed argument; the translation had to retain this factual, critical tone without smoothing it into promotional language.
- Examples and references such as The Federalist Papers or services like GPTZero and TurnItIn require contextual fidelity without adding explanatory material.

Quelltext	Übersetzung / Transcreation
AI Watermarking Won't Curb Disinformation	KI-Wasserzeichen sind kein wirksamer Schutz vor Desinformation
<i>By Jacob Hoffman-Andrews</i>	<i>Von Jacob Hoffman-Andrews</i>
Generative AI allows people to produce piles upon piles of images and words very quickly. It would be nice if there were some way to reliably distinguish AI-generated content from human-generated content. It would help people avoid endlessly arguing with bots online, or believing what a fake image purports to show. One common proposal is that big companies should incorporate watermarks into the outputs of their AIs. For instance, this could involve taking an image and subtly changing many pixels in a way that's undetectable to the eye but detectable to a computer program. Or it could involve swapping words for synonyms in a predictable way so that the meaning is unchanged, but a program could readily determine the text was generated by an AI.	Generative KI ermöglicht es, Unmengen von Bildern und Wörtern zu produzieren, und das in rasantem Tempo. Es wäre daher wünschenswert, wenn es eine Möglichkeit gäbe, KI-generierte Inhalte zuverlässig von Inhalten zu unterscheiden, die von Menschen erstellt wurden. So ließen sich endlose Online-Diskussionen mit Bots vermeiden und die Lügen in einem gefälschten Bild entlarven. Daher stand schon häufiger der Vorschlag im Raum, dass die großen Tech-Firmen Wasserzeichen in den Output ihrer KIs einbetten sollten. So könnten sie zum Beispiel Pixel eines Bildes subtil ändern – von einem menschlichen Auge nicht wahrnehmbar, für ein Computerprogramm jedoch leicht auszulesen. Bei Texten könnte der Algorithmus hingegen Wörter auf vorhersehbare Weise durch Synonyme ersetzen, sodass die Bedeutung unverändert bliebe, ein Programm jedoch mühelos feststellen könnte, ob der Text von einer KI generiert wurde oder nicht.
Unfortunately, watermarking schemes are unlikely to work. So far most have proven easy to remove, and it's likely that future schemes will have similar problems.	Leider dürften Wasserzeichen-Verfahren kaum funktionieren. Die bislang eingesetzten Methoden erweisen sich als leicht zu entfernen. Und zukünftigen Verfahren blüht wahrscheinlich das gleiche Schicksal.
One kind of watermark is already common for digital images. Stock image sites often overlay text on an image that renders it mostly useless for publication. This kind of watermark is visible and is slightly challenging to remove since it requires some photo editing skills.	Bei digitalen Bildern sind Wasserzeichen ja schon länger üblich: So legen Stockfoto-Anbieter oft Text über ein Bild, um es so für eine unlicenzierte Veröffentlichung unbrauchbar zu machen. Diese Art von Wasserzeichen ist sichtbar; es erfordert spezielle Bildbearbeitungskenntnisse, um sie zu entfernen:

Quelltext	Übersetzung / Transcreation
	
Images can also have metadata attached by a camera or image processing program, including information like the date, time, and location a photograph was taken, the camera settings, or the creator of an image. This metadata is unobtrusive but can be readily viewed with common programs. It's also easily removed from a file. For instance, social media sites often automatically remove metadata when people upload images, both to prevent people from accidentally revealing their location and simply to save storage space.	Kameras oder Bildbearbeitungsprogramme hängen oft auch Metadaten an Bilder an, zum Beispiel Informationen wie Datum, Uhrzeit und Ort der Aufnahme, die Kameraeinstellungen oder den Urheber eines Bildes. Diese Metadaten sind unauffällig, können aber mit gängigen Programmen mühelos ausgelesen werden. Doch ebenso leicht lassen sie sich auch entfernen: Zum Beispiel löschen Social-Media-Plattformen solche Metadaten oft automatisch aus hochgeladenen Bildern – sowohl um zu verhindern, dass versehentlich Standorte preisgegeben werden, als auch schlichtweg, um Speicherplatz zu sparen.
A useful watermark for AI images would need two properties:	Ein nützliches Wasserzeichen für KI-Bilder bräuchte zwei Eigenschaften:
It would need to continue to be detectable after an image is cropped, rotated, or edited in various ways (robustness).	Es müsste auch dann noch erkennbar sein, wenn ein Bild beschnitten, gedreht oder auf verschiedene Weise bearbeitet wurde (Robustheit).
It couldn't be conspicuous like the watermark on stock image samples, because the resulting images wouldn't be of much use to anybody.	Es dürfte nicht auffällig sein wie das Wasserzeichen auf Vorschaubildern von Stockfotos, weil die resultierenden Bilder für niemanden zu gebrauchen wären.
One simple technique is to manipulate the least perceptible bits of an image. For instance, to a human viewer these two squares are the same shade:	Eine einfache Technik besteht darin, die am wenigsten wahrnehmbaren Bits eines Bildes zu manipulieren. Für das menschliche Auge haben beispielsweise diese beiden Quadrate denselben Farbton:

Quelltext	Übersetzung / Transcreation
	
But to a computer it's obvious that they are different by a single bit: #93c47d vs 93c57d. Each pixel of an image is represented by a certain number of bits, and some of them make more of a perceptual difference than others. By manipulating those least-important bits, a watermarking program can create a pattern that viewers won't see, but a watermarking-detecting program will. If that pattern repeats across the whole image, the watermark is even robust to cropping. However, this method has one clear flaw: rotating or resizing the image is likely to accidentally destroy the watermark.	Für einen Computer ist jedoch offensichtlich, dass sie sich um ein einziges Bit unterscheiden: #93c47d vs. #93c57d. Jeder Pixel eines Bildes wird durch eine bestimmte Anzahl von Bits dargestellt; einige davon machen einen größeren wahrnehmbaren Unterschied aus als andere. Durch die Manipulation der am wenigsten bedeutsamen Bits kann ein Wasserzeichen-Programm ein Muster erzeugen, das nicht zu sehen, von einem Erkennungsprogramm jedoch leicht zu identifizieren ist. Wird dieses Muster über das gesamte Bild wiederholt, ist das Wasserzeichen sogar robust gegen das Beschneiden. Allerdings hat diese Methode eine offensichtliche Schwachstelle: Das Drehen oder Skalieren eines Bildes könnte das Wasserzeichen versehentlich zerstören.
There are more sophisticated watermarking proposals that are robust to a wider variety of common edits. However, proposals for AI watermarking must pass a tougher challenge. They must be robust against someone who knows about the watermark and wants to eliminate it. The person who wants to remove a watermark isn't limited to common edits, but can directly manipulate the image file. For instance, if a watermark is encoded in the least important bits of an image, someone could remove it by simply setting all the least important bits to 0, or to a random value (1 or 0), or to a value automatically predicted based on neighboring pixels. Just like adding a watermark, removing a watermark this way gives an image that looks basically identical to the original, at least to a human eye.	Es gibt ausgefeiltere Vorschläge für Wasserzeichen-Verfahren, die einer größeren Bandbreite gängiger Bearbeitungen widerstehen können. Vorschläge für KI-Wasserzeichen müssen jedoch eine höhere Hürde nehmen. Sie müssen auch resistent gegen das bewusste Entfernen sein: Wer ein Wasserzeichen löschen möchte, ist nicht auf gängige Bearbeitungen beschränkt, sondern kann die Bilddatei direkt manipulieren. Wenn ein Wasserzeichen zum Beispiel in den am wenigsten bedeutsamen Bits eines Bildes encodiert ist, ließe es sich entfernen, indem man einfach alle diese Bits auf 0, auf einen zufälligen Wert (1 oder 0) oder auf einen Wert setzt, der auf den benachbarten Pixeln basiert. Genau wie das Hinzufügen eines Wasserzeichens ergibt auch sein derartiges Entfernen ein Bild, das zumindest für das menschliche Auge vom Original praktisch nicht zu unterscheiden ist.
Coming at the problem from the opposite direction, some companies are working on ways to prove that an image came from a camera ("content authenticity"). Rather than marking AI generated images, they add metadata to camera-generated images, and use cryptographic signatures to prove the metadata is genuine. This approach is more workable than watermarking AI generated images, since there's no incentive to remove the mark. In fact, there's the opposite incentive: publishers would want to keep this metadata around because it helps establish that their images are "real." But it's still a fiendishly complicated scheme, since the chain of verifiability has to be preserved through all software used to edit photos. And most cameras will never produce this metadata, meaning that its absence can't be used to prove a photograph is fake.	Umgekehrt arbeiten einige Unternehmen an Nachweismöglichkeiten, dass ein Bild von einer Kamera stammt („Content Authenticity“ — Echtheitsnachweis für Aufnahmen). Anstatt KI-generierte Bilder zu kennzeichnen, fügen sie kameragenerierten Bildern kryptografisch signierte und somit als echt ausgewiesene Metadaten hinzu. Dieser Ansatz ist praktikabler als der Einsatz von Wasserzeichen in KI-generierten Bildern, denn es besteht ja kein Anreiz, diese Wasserzeichen zu entfernen – im Gegenteil: Publisher möchten diese Metadaten erhalten, weil sie damit nachweisen können, dass ihre Bilder „echt“ sind. Dennoch ist solch ein Verfahren ausgesprochen komplex, da die Nachweiskette über sämtliche zur Fotobearbeitung verwendete Software hinweg erhalten bleiben muss. Zudem erzeugen die meisten Kameras keine solchen Metadaten; ihr Fehlen ist also kein Beleg dafür, dass ein Foto eine Fälschung ist.

Quelltext			Übersetzung / Transcreation		
Comparing watermarking vs content authenticity, watermarking aims to identify or mark (some) fake images; content authenticity aims to identify or mark (some) real images. Neither approach is comprehensive, since most of the images on the Internet will have neither a watermark nor content authenticity metadata.			Vergleicht man KI-spezifische Wasserzeichen-Verfahren mit Content Authenticity, so zielen Wasserzeichen-Verfahren darauf ab, (einige) gefälschte Bilder zu identifizieren oder zu kennzeichnen; Content Authenticity erlaubt hingegen, (einige) echte Bilder zu identifizieren oder zu kennzeichnen. Keiner der beiden Ansätze greift umfassend, da die meisten Bilder im Internet weder ein Wasserzeichen noch Content-Authenticity-Metadaten aufweisen.		
	Watermarking	Content authenticity		Wasserzeichen-Verfahren	Content Authenticity
AI images	Marked	Unmarked	KI-Bilder	gekennzeichnet	nicht gekennzeichnet
(Some) camera images	Unmarked	Marked	(Einige) Kamerabilder	nicht gekennzeichnet	gekennzeichnet
Everything else	Unmarked	Unmarked	Alles andere	nicht gekennzeichnet	nicht gekennzeichnet
Text-based Watermarks			Wasserzeichen in Texten		
The watermarking problem is even harder for text-based generative AI. Similar techniques can be devised. For instance, an AI could boost the probability of certain words, giving itself a subtle textual style that would go unnoticed most of the time, but could be recognized by a program with access to the list of words. This would effectively be a computer version of determining the authorship of the twelve disputed essays in The Federalist Papers by analyzing Madison’s and Hamilton’s habitual word choices.			Das Problem der Wasserzeichen stellt sich bei textbasierter generativer KI sogar noch komplexer dar. Zwar sind vergleichbare Techniken denkbar: Zum Beispiel könnte eine KI die Wahrscheinlichkeit bestimmter Wörter erhöhen und sich so einen subtilen Textstil zulegen, der meistens unbemerkt bliebe, aber von einem Programm mit dem Zugriff auf die entsprechenden Wortlisten erkannt werden könnte. Dies wäre im Grunde eine Computerversion der Bestimmung der Autorschaft der zwölf umstrittenen Essays in The Federalist Papers durch die Analyse der typischen Wortwahl von Madison und Hamilton.		
But creating an indelible textual watermark is a much harder task than telling Hamilton from Madison, since the watermark must be robust to someone modifying the text trying to remove it. Any watermark based on word choice is likely to be defeated by some amount of rewording. That rewording could even be performed by an alternate AI, perhaps one that is less sophisticated than the one that generated the original text, but not subject to a watermarking requirement.			Das Erstellen eines unauslöschlichen Text-Wasserzeichens ist jedoch eine weitaus schwierigere Aufgabe, als Hamilton von Madison zu unterscheiden, da das Wasserzeichen der gezielten Manipulation des Textes widerstehen muss. Jedes auf Wortwahl basierende Wasserzeichen dürfte sich unwirksam machen lassen, indem man den Text ausreichend umformuliert. Diese Arbeit könnte sogar von einer alternativen KI vorgenommen werden – eine, die vielleicht weniger ausgefeilt sein mag als diejenige, die den Originaltext generiert hat, dafür aber keiner Wasserzeichenpflicht unterliegt.		
There’s also a problem of whether the tools to detect watermarked text are publicly available or are secret. Making detection tools publicly available gives an advantage to those who want to remove watermarking, because they can repeatedly edit their text or image until the detection tool gives an all clear. But keeping them a secret makes them dramatically less useful, because every detection request must be sent to whatever company produced the watermarking. That would potentially require people to share private communication if they			Es stellt sich zudem eine zentrale Frage: Sollen die Erkennungstools für Text mit Wasserzeichen öffentlich verfügbar sein oder geheim gehalten werden? Erkennungstools öffentlich verfügbar zu machen, macht das Entfernen des Wasserzeichens leichter, da man den Text oder das Bild so lange bearbeiten kann, bis das Erkennungstool „Entwarnung“ gibt. Diese Tools hingegen geheim zu halten, reduziert ihren Nutzen in anderer Hinsicht deutlich, da jede Erkennungsanfrage an den jeweiligen Anbieter des Wasserzeichenverfahrens gesendet		

Quelltext	Übersetzung / Transcreation
wanted to check for a watermark. And it would hinder attempts by social media companies to automatically label AI-generated content at scale, since they'd have to run every post past the big AI companies.	werden muss,. Unter Umständen hätte dies zur Folge, dass private Kommunikation zur Wasserzeichenprüfung an eine externe Partei übertragen würde. Zudem würde ein solcher Ansatz Versuche von Social-Media-Unternehmen behindern, KI-generierte Inhalte in großem Maßstab automatisch zu kennzeichnen, da sie jeden Beitrag von allen großen KI-Unternehmen mit ihren jeweils individuellen Wasserzeichen-Tools prüfen lassen müssten.
Since text output from current AIs isn't watermarked, services like GPTZero and TurnItIn have popped up, claiming to be able to detect AI-generated content anyhow. These detection tools are so inaccurate as to be dangerous, and have already led to false charges of plagiarism.	Aktuelle KIs kennzeichnen derzeit ihren Text-Output noch nicht Wasserzeichen. Aktuell überall aus dem Boden schießende Dienste wie GPTZero behaupten aber von sich, KI-generierte Inhalte trotzdem erkennen zu können. Diese Erkennungstools sind gefährlich ungenau – und haben schon zu einer Reihe vollkommen unbegründeter Plagiatsvorwürfe geführt.
Lastly, if AI watermarking is to prevent disinformation campaigns sponsored by states, it's important to keep in mind that those states can readily develop modern generative AI, and probably will in the near future. A state-sponsored disinformation campaign is unlikely to be so polite as to watermark its output.	Wenn KI-Wasserzeichen zudem von feindlich gesinnten Staaten gesteuerte Desinformationskampagnen verhindern sollen, gilt es schließlich zu bedenken, dass diese Staaten mühelos ihre eigene moderne generative KI entwickeln können und dies in naher Zukunft auch tun werden. Dabei werden sie im Umfeld von Desinformationskampagnen kaum so höflich sein, ihre Ausgaben mit Wasserzeichen zu versehen.
Watermarking of AI generated content is an easy-sounding fix for the thorny problem of disinformation. And watermarks may be useful in understanding reshared content where there is no deceptive intent. But research into adversarial watermarking for AI is just beginning, and while there's no strong reason to believe it will succeed, there are some good reasons to believe it will ultimately fail.	Wasserzeichen in KI-generierten Inhalten sind nur scheinbar eine einfache Lösung für die komplexen Herausforderungen der Desinformation. Wasserzeichen mögen sinnvoll sein, um den KI-Einsatz in ohne Täuschungsabsicht weiterverbreiteten Inhalten kenntlich zu machen. Doch die Forschung zu adversarialen Wasserzeichenverfahren für KI steht zwar erst am Anfang, allerdings spricht wenig für ihren Erfolg – und viel für ihr Scheitern.

Impressum / Imprint

Übersetzung von / Transcreation by

Helmut Barz

Blumenstraße 52

D-63069 Offenbach am Main

helmut@helmut-barz.com

Hinweis / Note

Dieser Text entstand als Arbeitsprobe auf eigene Initiative und ohne Auftraggeber. Weitere Arbeitsproben finden Sie unter: <https://helmut-barz.com/arbeitsproben/>

This text was written as a sample piece on my own initiative and without a client. You can find more about my service offerings (including additional samples) at:
<https://translations.helmut-barz.com>

Rechte / Rights:

Dieses Dokument ist unter der Creative-Commons-Lizenz „Namensnennung 3.0 Vereinigte Staaten“ (CC BY 3.0 US) lizenziert. Das bedeutet: Sie dürfen das Material vervielfältigen, verbreiten und bearbeiten (auch kommerziell), solange Sie die Urheberschaft angemessen nennen und einen Link zur Lizenz beifügen.

Vollständiger Lizenztext: <https://creativecommons.org/licenses/by/3.0/us/deed.de>

This document is licensed under the Creative Commons Attribution 3.0 United States (CC BY 3.0 US) License. This means you are free to copy, distribute, and adapt the material (even commercially), provided you give appropriate credit and provide a link to the license.

Full license text: <https://creativecommons.org/licenses/by/3.0/us/>