TAGESSPIEGEL
BACKGROUND

Digitalisierung & KI

Mit Wasserzeichen gegen Deepfakes?

In der Politik wächst die Sorge vor Desinformation durch Deepfakes. Microsoft hat nun eine Wasserzeichen-Technologie eingeführt, um dem Problem zu begegnen. Doch damit diese wirklich etwas bewegen kann, müsste sie sich als Standard etablieren.

von **Kilian Schroeder** veröffentlicht am 09.01.2024

Olaf Scholz steht im Kanzleramt, der Reichstag ist durch das Fenster im Hintergrund zu sehen und links von ihm steht die deutsche Fahne: „Liebe Mitbürgerinnen und Mitbürger, ich wende mich heute an Sie, weil unser Land einer schweren Bedrohung ausgesetzt ist“, sagt der Kanzler. Er sagt auch, er wolle die AfD verbieten lassen. Nur: Weder hat die Bundesregierung so etwas vor, noch hat Scholz je eine solche Rede gehalten. Das Video ist ein **Deepfake**, Teil eines Projektes des **Zentrums für politische Schönheit**.

Zwar sind einige Details im Video, zum Beispiel die **Stimme, schon verräterisch**. Doch die Bundesregierung war von der Aktion des selbsternannten Künstler:innenkollektivs alarmiert: „Solche Deepfakes sind **kein Spaß**“, postete Regierungssprecher **Steffen Hebestreit** kurz nach der Veröffentlichung. Denn das Video ist nur eines von vielen Beispielen, die zeigen, dass die Deepfake-Technologie Einzug in die Politik hält. Und die Sorge wächst, dass feindliche Akteure sie ausnutzen könnten, um falsche Videos und Bilder von Politiker:innen zu verbreiten, die irgendwann niemand mehr als Fälschung erkennt.

Kryptografische Methoden sollen Deepfakes entlarven

Microsoft will dieser Gefahr nun mit sogenannten „**Content Credentials**“ begegnen. Im Oktober vergangenen Jahres führte der Konzern die „Icons of Transparency“ für alle KI-generierten Bilder ein, die mit Bing Image Creator erstellt wurden. Die Idee: kryptografische Wasserzeichen. Diese sollen **Metadaten** enthalten, an denen die **Echtheit der Bilder und Videos erkennbar** ist. Wenn Nutzer:innen über diese Wasserzeichen scrollen, sollen sie zukünftig unter anderem sehen können, von wem das Video ist, wann und wie es erstellt wurde – und ob KI dahinter steckt.

Deepfakes ließen sich leichter enttarnen, weil die Content Credentials **fester Bestandteil der Geschichte des Bildes oder Videos** seien, argumentiert Microsoft. Nutzer:innen könnten also sehen, ob ein Post im Nachhinein digital verändert wurde.

Zunächst bietet Microsoft die Technologie nur **politischen Kampagnen für die US-Präsidentschaftswahlen** an. „Im Jahr 2024 könnte die Welt erleben, dass mehrere autoritäre Staaten versuchen, in Wahlprozesse einzugreifen“, beschreibt **Brad Smith**, Präsident von Microsoft, in der Ankündigung des neuen Tools. Microsoft wolle Kandidat:innen, Kampagnen und Wähler:innen helfen, „mehr Kontrolle über ihre Inhalte und ihr Erscheinungsbild zu behalten“, antwortet der Technologiekonzern auf Anfrage von Tagesspiegel Background.

KI-generierte Bilder aus dem **Bing Image Creator** würden nun sowohl mit einem Bing-Logo als Wasserzeichen, als auch mit den kryptografischen Content Credentials versehen. Derzeit gebe es die Services erst in einer privaten Vorschau, im Frühjahr sollen sie den Wahlkampagnen in den USA zur Verfügung gestellt werden.

Politik macht den Konzernen Druck

Neben der plakativ geäußerten Fürsorge zum Schutz der Demokratie dürften auch **wirtschaftliche Motive** hinter der Ankündigung des Computer- und Softwarekonzerns stehen. Im Oktober beauftragte US-Präsident **Joe Biden** *in einer „Executive Order“ (https://background.tagesspiegel.de/digitalisierung/usa-untermauern-per-dekret-globalen-ki-fuehrungsanspruch)* das Handelsministerium,

Richtlinien für eine Wasserzeichen-Technologie zu entwickeln. Schon zuvor hatten mehrere Senator:innen beider Parteien ein Gesetzesvorhaben im Kongress eingebracht, das es politischen Kampagnen sogar verbieten soll, irreführende KI-generierte Inhalte zu verbreiten („Protect Elections from Deceptive AI Act“). Microsoft unterstützt das Vorhaben laut eigener Aussage, auch **Meta plant schärfere Kontrollen** von KI-generierten Inhalten.

Ob und wann Microsoft seine Krypto-Wasserzeichen in Deutschland zugänglich machen will, ist unklar. Dabei geht die Sorge vor Deepfakes in der Politik nicht erst seit dem oben genannten Video um. Innenministerin **Nancy Faeser** (SPD) hatte dem „Handelsblatt“ bereits im September gesagt, dass Deepfakes ein „sehr gefährliches Mittel“ seien, um öffentliche Debatten zu manipulieren. Das **Bundesamt für Sicherheit in der Informationstechnik** (BSI) warnt auf seiner Website vor Desinformationskampagnen mit gefälschten Inhalten. Laut einer Umfrage des IT-Branchenverbandes Bitkom sehen etwa 60 Prozent der Menschen hierzulande in Deepfakes eine Gefahr für die Demokratie.

Zwar wurde auch in Brüssel um eine **KI-Kennzeichnungspflicht** im kürzlich beschlossenen AI Act gerungen (Tagesspiegel Background *berichtete (https://background.tagesspiegel.de/digitalisierung/bruessel-ringt-um-ki-kennzeichnungspflicht)*). Die Verantwortung zur Kennzeichnung KI-generierter Inhalte wollen die Gesetzgeber aber eher beim **Endnutzer verorten**, eine verpflichtende, softwareseitige Markierung in Form von Wasserzeichen ist bislang nicht vorgesehen. Von der Vorschrift für Endnutzer, offenzulegen, dass die Inhalte künstlich erzeugt oder manipuliert wurden, werden sich die viel zitierten böswilligen Akteure wohl kaum beeindrucken lassen.

Mit Blick auf die Europawahl forderte die EU-Kommission die politischen Parteien und Kampagnenorganisationen immerhin in einer gezielten *Empfehlung (https://commission.europa.eu/system/files/2023-12/C_2023_8626_1_EN_ACT_part1_v5.pdf)* auf, auf die Nutzung KI-generierter Deepfakes in ihren Kampagnen zu verzichten. Die Parteien sollten sich hier am besten auf Codes of Conducts einigen. Zudem wird

gefordert, dass der Einsatz von KI-Systemen für politische Wahlwerbung transparent gemacht wird.

Wirkung nicht überschätzen

Welche Schlagkraft haben Werkzeuge wie das von Microsoft entwickelte, um Gesellschaften vor gefährlichen Deepfakes zu schützen? **Christian Hoffmann**, Kommunikationswissenschaftler an der Uni Leipzig, rät dazu, die Wirkung dieser **Wasserzeichen nicht zu überschätzen**. „Ich bin mir nicht sicher, ob das etwas hilft", sagt Hoffmann, der mit seinem „Deepfake Project" verstärkt zur gesellschaftlichen Wirkung solcher manipulativen Fälschungen forscht.

Das größte Problem: Damit die Content Credentials vor falschen Informationen schützen, müssten die Nutzer:innen verstehen, was sie da sehen. „Es müsste sich erst einmal **als Standard etablieren**." Damit das klappt, müssten sich die großen Konzerne **auf eine solche Technologie einigen**.

Doch Hoffmann sieht auch Stärken der Technologie: Sie könne **Journalist:innen helfen**, einfacher falsche Videos zu entdecken. „Die Frage ist: Wie können wir sicherstellen, dass Journalistinnen und Journalisten nicht auf Deepfakes hereinfallen? Und das scheint mir der Charme von diesen Wasserzeichen zu sein", sagt Hoffmann. Wenn Journalist:innen Deepfakes leichter enttarnten, könne das **dem öffentlichen Diskurs zugutekommen**. *Kilian Schroeder*