

## / Policy Brief

# Ergänzung in der KI-Verordnung: Erweiterung des Artikels 5 über „Verbotene Praktiken im KI-Bereich“

April 2026

## Zusammenfassung

Wir begrüßen die geplante Ergänzung von Artikel 5 der KI-Verordnung um ein Verbot nicht-einvernehmlicher sexualisierter Deepfakes. Sie wäre ein wichtiger Baustein zum Schutz Betroffener digitaler sexualisierter Gewalt. Die Evidenz ist eindeutig: Solche Inhalte betreffen ganz überwiegend Frauen, stellen eine Form geschlechtsspezifischer Gewalt dar und entfalten einen „Silencing-Effekt“, der demokratische Teilhabe einschränkt. Ein solches Verbot muss jedoch gezielt ausgestaltet sein, damit es legitime Nutzungen nicht erfasst und Haftung klar zugeordnet wird.

Unsere Forderungen im Überblick:

- **Einverständnis präzise definieren:** Frei, informiert, kontextbezogen, ausdrücklich; keine Erfassung nicht-realer Personen.
- **Haftungszuordnung eindeutig regeln:** Nur KI-Systeme ohne angemessene Schutzmaßnahmen erfassen.
- **Unbeabsichtigte Nebeneffekte vermeiden:** Open-Source-KI-Entwicklung nicht durch Überregulierung gefährden.
- **Sicherheitsmaßnahmen stärken:** Angemessene Schutzvorkehrungen mit kontinuierlicher Überprüfung und Meldemöglichkeiten.
- **Einverständnisabfrage:** Bei Darstellungen realer Personen verpflichtend machen.

## Kontext, Evidenz und Forderungen

Die KI-Verordnung (KI-VO) der EU wird derzeit überarbeitet (sog. KI-Omnibus-Verfahren). Die Vorschläge des [Europäischen Parlaments](#) (EP) und des [Rats der Europäischen Union](#) (Rat) sehen eine Ergänzung von Artikel 5 der KI-VO vor, der verbotene Praktiken definiert. Die Erstellung von nicht-einvernehmlichen sexualisierten Deepfakes soll in diesem Abschnitt als weiterer Anwendungsfall aufgenommen werden.

Wir sprechen uns deutlich für diese Ergänzung aus. Bestehenden Regelungen wie der Digital Services Act (DSA) adressieren vor allem, wie solche Inhalte auf Plattformen verbreitet werden. Nationale Vorhaben, wie das Gesetz gegen Digitale Gewalt, das gerade entwickelt wird, zielen darauf ab, das Strafrecht durch neue Tatbestände zu aktualisieren. Eine Ergänzung in der KI-VO würde eine wichtige regulatorische Lücke schließen, indem sie gezielt Anbieter und Betreiber von KI-Systemen in die Pflicht nimmt, und wäre damit ein weiterer wichtiger Baustein zum Schutz von Betroffenen digitaler sexualisierter Gewalt.

### Problemstellung

Mit der zunehmenden Verbreitung von KI, insbesondere KI mit allgemeinem Verwendungszweck (General Purpose AI, GPAI), wird es deutlich einfacher, Bilder, Videos und Tonaufnahmen zu erstellen oder zu manipulieren. Problematisch sind dabei solche KI-Modelle und auf ihnen aufbauende Tools, die es erleichtern, nicht-einvernehmliche sexualisierte Deepfakes zu erstellen. Das Spektrum reicht hier von Anwendungen, die ausdrücklich damit werben, Bilder realer Personen sexualisieren zu können (meist, ohne nachzuweisen, dass Einvernehmlichkeit herrscht), bis hin zu allgemeiner Software, wie „Face-Swapping“-Apps, die für solche Zwecke missbraucht werden können.

Die meisten großen GPAI-Systeme (wie ChatGPT, Claude etc.) verfügen bereits über Sicherheitsvorkehrungen, damit derartige Inhalte nicht erzeugt werden können. Auch Open-Source-KI kann entsprechende Schutzmaßnahmen beinhalten. Jedoch ist es technisch nahezu unmöglich, mit absoluter Sicherheit zu garantieren, dass Sicherheitsmaßnahmen nicht ausgehebelt werden können. Nutzer\*innen diskutieren, beispielsweise in einschlägigen Foren, neue Wege, wie sie Sicherheitsvorkehrungen umgehen können. Einen Anbieter oder Betreiber für jede denkbare Nutzung seines Produkts haftbar zu machen, bringt daher erhebliche Probleme mit sich.

### Forderungen

Ein ergänzendes Verbot in der KI-VO muss so ausgestaltet sein, dass es legitime Arten, Nacktdarstellungen zu erzeugen oder zu bearbeiten, nicht erfasst, etwa wenn keine realen Personen betroffen oder Darstellungen einvernehmlich entstanden sind.

Dennoch gibt es Technologien, die es eindeutig erleichtern, nicht-einvernehmliche sexualisierte Deepfakes zu erzeugen – und genau diese Anbieter und Betreiber müssen in Verantwortung genommen werden. Da eine solche Verantwortung auch mit entsprechenden Strafen einhergeht, ist hier zu differenzieren: Das Gesetz soll nicht alle KI-Systeme zur Bilderstellung verbieten, sondern diejenigen, die es ohne wirksame Schutzvorkehrungen zulassen, derartige Inhalte zu erzeugen. Die Sicherheitsmaßnahmen sollten so streng wie möglich gestaltet sein und müssen regelmäßig von Anbietern oder Betreibern geprüft und getestet werden.

Zwar kann es auch erheblichen Schaden für Betroffene verursachen, wenn Sicherheitsvorkehrungen vorübergehend umgangen werden. Haften sollten jedoch nur Anbieter und Betreiber von KI-Systemen, die keine *angemessenen* Schutzmaßnahmen getroffen haben. Systeme, deren Maßnahmen erst mit erheblichem Aufwand umgangen wurden, sollten nicht pauschal verboten werden, da eben eine Deaktivierung von Nutzungsbeschränkungen (Jailbreak) technisch nie hundertprozentig auszuschließen ist. Wird ein KI-Modell von Dritten als Grundlage genutzt, um darauf aufbauend eigene Anwendungen zu entwickeln, und werden dabei bestehende Sicherheitsvorkehrungen entfernt oder umgangen, sollte nicht der Anbieter des zugrundeliegenden Modells haftbar gemacht werden, sondern diejenigen, die dessen Schutzmaßnahmen beseitigt und das Modell zweckentfremdet haben. In solchen Fällen sollte das neu entstandene System eigenständig bewertet werden.

Im Folgenden heben wir – auch mit Blick auf die teils bereits vorliegenden Formulierungen in den Vorschlägen des EP und des Rates – die aus unserer Sicht besonders relevanten Aspekte hervor:

- **Definition von Einverständnis:** Eine präzise Definition des Einverständnisses ist für den Bereich der Erstellung sexualisierter Darstellungen unverzichtbar. Die Zustimmung muss frei von Zwang, kontextbezogen, informiert, unmissverständlich und ausdrücklich erfolgen, damit einvernehmlich erstellte Inhalte klar von nicht-einvernehmlich erstellten Inhalten unterschieden werden können. Dabei ist klarzustellen, dass ein Verbot grundsätzlich nicht die Erstellung von Darstellungen nicht-realer Personen umfassen sollte.
- **Klarheit bei der Haftungszuordnung:** Es muss eindeutig geregelt sein, unter welchen Umständen KI-Systeme zur Bild-, Video- und Tonbearbeitung unter das Verbot in der KI-VO fallen. Es sollten nur KI-Systeme von der Regelung erfasst werden, die es erleichtern, nicht-einvernehmliche sexualisierte Deepfakes zu erstellen *und* dabei weder *angemessene* Schutzvorkehrungen treffen noch bei einem Missbrauch Korrekturmaßnahmen ergreifen. Maßgeblich hierfür sollten dabei die Funktionalitäten, Anwendungsmöglichkeiten, Zielsetzung und Trainingsdaten des jeweiligen Systems sein.
- **Unbeabsichtigten Nebeneffekten entgegenwirken:** Ein Verbot in Artikel 5 der KI-VO könnte insbesondere im Bereich von Open-Source-KI über das Regulierungsziel hinauschießen: Anbieter und Betreiber, die angemessene Sicherheits- und Korrekturmaßnahmen treffen und deren Systeme die Erstellung nicht-einvernehmlicher sexualisierter Deepfakes nicht explizit erleichtern, sollten nicht für eine Zweckentfremdung durch Dritte oder einen Jailbreak haftbar gemacht werden. Daher ist eine eindeutige Haftungszuordnung, wie im vorangegangenen Punkt dargelegt, unerlässlich. Darüber hinaus sollte geprüft werden, welche ergänzenden Maßnahmen geeignet wären, (Open-Source-)Anbieter und -Betreiber, die angemessene Schutzvorkehrungen treffen, vor den Folgen eines Missbrauchs ihrer Systeme zu schützen.
- **Sicherheitsmaßnahmen vorsehen:** KI-Systeme und GPAI-Modelle, die dazu genutzt werden können, zur Bild-, Video- und Tonaufnahmen zu erstellen, müssen sowohl auf Ebene ihrer Grundlagenmodelle als auch über die Eingabemöglichkeiten (Prompting) sicherstellen, dass wirksam verhindert wird, nicht-einvernehmliche sexualisierte Darstellungen zu erzeugen. Schutzmaßnahmen müssen kontinuierlich überprüft und gegebenenfalls erweitert werden, um ihre Umgehung zu erschweren. Anwender\*innen solcher Systeme

muss zudem die Möglichkeit gegeben werden, potenzielle Schutzlücken an die Anbieter oder Betreiber zu melden.

- **Verpflichtende Einverständnisabfrage bei Darstellungen realer Personen:** Apps und GPAI, die es erlauben, Bild-, Video- und Tonaufnahmen zu bearbeiten, müssen – außerhalb der durch Kunstfreiheit und Satire gedeckten Anwendungsbereiche – das Einverständnis dargestellter Personen abfragen, wenn Inhalte erstellt werden, die reale Personen zeigen. Darüber hinaus sollte transparent sein, welche KI-Modelle den jeweiligen Anwendungen zugrunde liegen.

## Unsere Organisation

**AlgorithmWatch** ist eine gemeinnützige Nichtregierungsorganisation mit Sitz in Berlin und Zürich. Wir setzen uns dafür ein, dass Algorithmen und Künstliche Intelligenz Gerechtigkeit, Demokratie, Menschenrechte und Nachhaltigkeit stärken, statt sie zu schwächen. Unsere Vision ist eine Welt, in der Technologie im Allgemeinen und algorithmische Systeme im Besonderen den Menschen zugutekommen. Die Systeme sollen Gesellschaften gerechter, demokratischer, inklusiver und nachhaltiger machen – sei es hinsichtlich zugeschriebener Herkunft und Gender, Rassifizierung, sexueller Orientierung, Alter, Klasse und Wohlstand oder Ressourcenverbrauch.