

Aktuelles aus der Spruchpraxis: Internetseite mit KI-Text-zu-Bild-Generator & KI-Chatbot ermöglicht Erstellung und Verbreitung von Darstellungen sexueller Gewalt an Minderjährigen

KI ist im Alltag von Kindern und Jugendlichen angekommen: Rund 91,2 Prozent der Jugendlichen (JIM-Studie 2025 S. 64) und neun Prozent der Kinder (KIM-Studie 2024, S. 35) nutzen Anwendungen künstlicher Intelligenz (KI) im Alltag. Besonders beliebt sind kostenfreie und unbeschränkte Text-zu-Bild-Generatoren. Im Dezember 2025 hat die Prüfstelle für jugendgefährdende Medien über ein Angebot entschieden, welches kostenfrei und in unbegrenztem Umfang die Bilderstellung mittels künstlicher Intelligenz ermöglichte.

Die Domain-Unterseite wurde in den nicht-öffentlichen Teil der Liste jugendgefährdender Medien aufgenommen. Das Gremium begründete die Listenaufnahme damit, dass das Medium kinder- und jugendpornografische Medieninhalte (auch CSAM [child sexual abuse material] genannt) innerhalb Foren- und Galeriebereichen bereithielt, deren Erzeugung ermöglichte und damit schwer jugendgefährdend wirkte.

Das englischsprachige Angebot war über verschiedene, genutzte Open-Source-Text-zu-Bild-Modelle und Künstliche-Intelligenz-Technologien in der Lage – je nach gewünschter Stilrichtung – zum Beispiel fotorealistische Bilder aus beliebigen Texteingaben zu erzeugen. Nutzende konnten aus über 70 verschiedenen Stilrichtungen wählen. Zudem verfügte das Angebot über weitere Funktionen, welche die Erstellung einer Aufgabenstellung (sogenannter Prompt) unterstützte oder gänzlich übernahm. Der Generator konnte kostenfrei, in zeitlich und mengenmäßig unbegrenzter Weise und ohne vorherigen Registrierungsvorgang genutzt werden. Rezipierende konnten Bilder durch die Eingabe von Prompts

innerhalb der Benutzeroberfläche des verfahrensgegenständlichen Mediums vornehmen. Dabei war etwa die Erzeugung unter anderem auch von Bildern, wie etwa einem Fantasiewesen, welches sich aus einem Brotlaib und einem Hamster zusammensetzt oder freizügigen, sexualisierten Darstellungen von erwachsenen Personen möglich. Neben der Möglichkeit nicht soziaethisch desorientierendes Material zu erstellen, war es auch möglich, kinder- und jugendpornografischem Material zu realisieren. Die erstellten Bilder konnten als Grundlage für die Generierung eines Chatbots mit anpassbarem Charakter genutzt werden. Je nach Ausrichtung des Charakters beziehungsweise des sich anschließenden Dialogs war es möglich, den Chatbot zur Ausgabe textlicher Kinderpornografie zu manipulieren. Zudem fand sich auf dem zu prüfenden Medium ein Forenbereich, in welchem unter anderem der Austausch über geschicktes Prompting zur Umgehung von Sicherungsmechanismen der geprüften Plattform diskutiert wurde. Auch wurden dort Prompts geteilt, welche die Erstellung von virtuellen Missbrauchsdarstellungen an Kindern und Jugendlichen ermöglichten sowie Hinweise auf Galeriebereiche verbreitet, welche solche Inhalte zeigten. Nach wenigen Nutzungen des Text-zu-Bild-Generators konnte ein Foren- und Galeriebereich eingesehen werden, welcher mittels einer Stichwortsuche nach Themengebieten durchsucht werden konnte.

Das Medium wies eine Sicherheitsarchitektur in Form der Nichtannahme bestimmter Prompts (sogenannte Prompt-Blacklist) sowie das Verhindern der Ausgabe von schädlichen Inhalten (Diskriminatoren) auf, jedoch hielt das verfahrensgegenständliche

Medium selbst innerhalb seines Forenbereichs leicht auffindbare Umgehungsmöglichkeiten dieser Sicherungsmechanismen vor.

Bei Eingabe von codierten Begriffen, welche in der pädokriminellen Szene Verwendung finden, konnten Forenbereiche aufgefunden werden, in welchen sich Nutzende über die konkrete Formulierungsweise von Prompts austauschten, sodass die Sicherungsmechanismen des Mediums umgangen und so Missbrauchsdarstellungen erstellt werden konnten. Darüber hinaus wurden dort Bezeichnungen von Galerien geteilt, welche bei Nutzung der Galeriesuchfunktion in der Ausgabe von Galerien mit Missbrauchsdarstellungen an Kindern und Jugendlichen resultierten.

Austausch und Verbreitung von Missbrauchsdarstellungen in versteckten Foren- und Galeriebereichen & Selbst-Generierungsmöglichkeit

Nach Ansicht des Gremiums sei der Austausch über Prompts, welche dazu geeignet sind, die Sicherungsmechanismen des Text-zu-Bild-Generators zu umgehen, um Missbrauchsdarstellungen erstellen zu können, als im höchsten Maße bedenklich zu bewerten. Kinder und Jugendliche könnten die Möglichkeit, existierende Beschränkungen zu umgehen, als Herausforderung ansehen. Dies führe nicht nur zur Erstellung weiteren Missbrauchsmaterials, sondern reize Minderjährige dazu an, eigene Umgehungsmöglichkeiten zu finden. Zudem könnten Minderjährige das erfolgreiche Überwinden von Sicherheitsmaßnahmen als Belohnung empfinden.

Es bestehe des Weiteren die erhebliche Gefahr, dass Kinder und Jugendliche die Darstellung virtueller, sexueller Handlungen an Minderjährigen als alltäglich oder gar legitim betrachten. Zudem sei es Minderjährigen durch das Einsehen der zur Erstellung der entsprechenden Darstellungen verwendeten Prompts möglich, diese selbst in identischer oder individuell abgewandelter Form bei eigenen Generierungen zu nutzen.

Insgesamt sei der niedrigschwellige Zugang zu Missbrauchsdarstellungen von Kindern und Jugendlichen im höchsten Maße als sozialetisch desorientierend zu betrachten.

Interaktionsmöglichkeit mit CSAM-Chatbots (Simulierung von Sexting & Cybergrooming)

Als besonders schwerwiegenden Aspekt des Mediums sah das Gremium die Erstellung von

KI-Chatbots an, welche auf zuvor generierten Missbrauchsdarstellungen von Kindern und Jugendlichen beruhte. Die Chatbots könnten mangels hinreichender Inhaltsmoderation und Schutzmechanismen zum Sexting mit einer virtuellen, minderjährigen Person genutzt werden. Insbesondere vor dem Hintergrund, dass durch geschicktes Prompting die anfänglich ablehnende Haltung der KI-Chatperson bezüglich sexueller Inhalte überwunden werden könne, sah das Gremium als Möglichkeit zum Trainieren von Cybergrooming mit realen Minderjährigen an.

Kritik der Nutzungsweise durch Nutzendengemeinschaft

Bei seiner Entscheidung berücksichtigte das Gremium Darstellungen sowie Äußerungen im Galerie- und Forenbereich, welche die Nutzung des KI-Modells zur Erstellung von Darstellungen sexueller Gewalt kritisierten. So ließen sich etwa die folgenden Botschaften in Text- beziehungsweise Bildform finden: „stop posting child p0rn“ oder „KILL PEDOS“. Diese seien jedoch nicht dazu geeignet, eine ausreichend distanzschaffende Wirkung zu den schwer jugendgefährdenden Inhalten zu erzeugen.

Abwägung mit den Grundrechten der Anbieter und dem Verhältnismäßigkeitsgrundsatz

Nach Ansicht des Gremiums überwogen die Belange des Jugendschutzes die Meinungs- und Kunstfreiheit der Anbieter. Zwar enthalte das Internetangebot eine Vielzahl unverfänglicher Darstellungen, diese stünden jedoch den virtuellen Darstellungen sexueller Gewalt gegenüber. Es bestehe ein vom Umfang nicht abschätzbares Potential, entsprechende Inhalte mit Hilfe des verfahrensgegenständlichen Angebots zu erstellen. Die im Medium selbst verbreiteten Befehle stellen dabei nur einen Bruchteil der potentiellen Umgehungsmöglichkeiten der implementierten Inhaltsmoderation dar. Es bestehe die Gefahr, dass Kinder und Jugendliche die Erstellungsmöglichkeit als niedrigschwelligen Einstieg in entsprechende sexuelle Tendenzen wahrnehmen und dass Desensibilisierungs- und Normalisierungseffekte durch die Wahrnehmung sexualisierter Gewalt gegenüber Minderjährigen eintreten.