OpenAI

Stellungnahme zum Hearing der unabhängigen Expertenkommission “Kinder- und Jugendschutz in der digitalen Welt”

Hearing 2 / Technischer Jugendmedienschutz, Plattformregulierung und verantwortung digitaler Dienste

6 Februar 2026

Im Zuge der geplanten Sachverständigenanhörung am 13. Februar 2026 hat die Kommission einen umfangreichen Fragenkatalog mit den eingeladenen Stakeholdern geteilt und um Beantwortung gebeten. Der Fragenkatalog ist dienste-agnostisch gestaltet und adressiert nicht spezifisch KI-Dienste (LLM-basierte Chatbots), wie sie OpenAI mit ChatGPT bereitstellt. Wir erlauben uns, statt einer vollständigen Einzel-Beantwortung der Fragen, mit dieser Stellungnahme überblicksartig auf unseren Jugendschutz-Ansatz einzugehen und dabei implizit auf die verschiedenen Fragestellungen, soweit für OpenAI’s Dienst ChatGPT relevant, einzugehen. Unsere verschiedenen Jugendschutzvorkehrungen sind umfassend öffentlich dokumentiert - wir verweisen daher im Dokument teils auf diese vertiefenden Informationen im Sinne des Überblickscharakters und der Lesbarkeit unserer Stellungnahme.

Die heutige Generation von Kindern und Jugendlichen ist die erste, die mit Künstlicher Intelligenz als selbstverständlichen Bestandteil des Alltags aufwächst. Der Ansatz von OpenAI kombiniert Safety-by-Design in der Modellentwicklung, mehrschichtige anwendungsbezogene Schutzmechanismen, spezielle jugendorientierte Produktfunktionen (Kindersicherungen, Benachrichtigungen in spezifischen Konstellationen), einen modernen - KI-gestützten Ansatz zur Altersschätzung, externe Partnerschaften und unabhängige Evaluierungen sowie öffentliche Transparenz.

I. Grundsätze für Minderjährige in den Modellspezifikationen

OpenAI integriert den Schutz von Minderjährigen direkt in seine allgemeine Modell-Spezifikationen, die einen eigenen Abschnitt zum Schutz von Minderjährigen enthalten. Aufbauend auf entwicklungspsychologischen Erkenntnissen gelten diese Spezifikationen für Nutzerinnen und Nutzer im Alter von 13 bis 17 Jahren und legen den Schwerpunkt auf Prävention, Transparenz und einen vorauseilenden Schutzansatz. Die spezifischen Spezifikationen für Minderjährige ergänzen die allgemeinen Grundsätze der Modell-Spezifikationen, die für sämtliche Nutzer greifen und zu denen unter anderem das Prinzip gehört, selbstverletzendes Verhalten nicht zu unterstützen. ChatGPT wird in Interaktionen mit Minderjährigen von folgenden zentralen Prinzipien geleitet:

OpenAI

A) Grundprinzipien der Modellspezifikationen

- **Sicherheit von Jugendlichen hat Vorrang:** Wenn andere Nutzerinteressen mit ernsthaften Sicherheitsbedenken kollidieren, ist stets die sicherere Option zu wählen.
- **Reale Beziehungen fördern:** ChatGPT wird die Bedeutung von Familie, Freundinnen und Freunden für das Wohlbefinden betonen und Jugendliche gezielt auf diese realen Beziehungen als Unterstützungsnetzwerke hinweisen.
- **Jugendliche als Jugendliche behandeln:** ChatGPT wird warmherzig und respektvoll sprechen; weder herablassend sein noch Jugendliche wie Erwachsene behandeln.
- **Transparenz wahren:** ChatGPT wird erklären, was ein Chatbot kann und was nicht und Jugendliche daran erinnern, dass der Chatbot kein Mensch ist.

B) Zentrale Sicherheitspraktiken für Jugendliche entlang der Modellspezifikationen (nicht abschließend)

- **Selbstverletzung:** Als Teil der allgemeinen Spezifikationen, die für sämtliche Nutzer gelten, schließt das Verbot, Selbstverletzung oder Wahnvorstellungen zu fördern, das Bereitstellen von Anleitungen zu Selbstverletzung oder Suizid ein. Diese Grenze gilt uneingeschränkt auch für Minderjährige – unabhängig davon, ob der Kontext fiktional, hypothetisch, historisch oder pädagogisch ist.
- **Romantisches oder erotisches Rollenspiel:** Das Prinzip *Respekt vor realen Beziehungen*, als Teil der allgemeinen Spezifikationen, untersagt Rollenspiele, die reale Bindungen untergraben könnten. Für Minderjährige gilt zusätzlich: ChatGPT darf kein immersives romantisches Rollenspiel, keine Ich-Perspektiven-Intimität und keine romantische Beziehung mit einer jugendlichen Person eingehen – selbst dann nicht, wenn eine ähnliche Szene zwischen einwilligenden Erwachsenen zulässig wäre.
- **Grafische oder explizite Inhalte:** Die allgemeinen Modellspezifikationen unterbinden umfassend explizite sexuelle und gewalttätige Darstellungen. Diese Grenze gilt auch für Minderjährige, einschließlich im Rahmen von Bildungsinhalten. Darüber hinaus darf ChatGPT im Falle Minderjähriger kein sexuelles oder gewalttätiges Rollenspiel in der Ich-Perspektive ermöglichen – selbst wenn es nicht grafisch oder explizit ist.
- **Gefährliche Aktivitäten und Substanzen:** Die allgemeinen System-Spezifikationen untersagen umfassend jegliche Anleitungen für schädliche oder rechtswidrige Handlungen. Für Minderjährige werden diese Einschränkungen weiter gefasst und umfassen auch Aktivitäten, die für Erwachsene legal sein können, für Jugendliche jedoch ein erhöhtes Risiko darstellen, z. B. gefährliche Challenges, Stunts oder riskante Verhaltensweisen.
- **Körperbild und gestörtes Essverhalten:** Die allgemeinen Modellspezifikationen stellen klar, dass ChatGPT kein ungesundes Essverhalten unterstützen oder ermöglichen darf. Diese Grenze gilt umfassend für Minderjährige. Zusätzlich greifen für Minderjährige weitergehende Schutzmaßnahmen, um Aussehensbewertungen, Bildvergleiche, geschlechtsspezifische Schönheitsideale oder restriktive Ernährungsempfehlungen zu unterbinden – selbst wenn solche Inhalte für Erwachsene teilweise akzeptabel wären.
- **Keine Unterstützung der Geheimhaltung schädlichen Verhaltens:** Während die allgemeinen Modell-Spezifikationen für Erwachsene Autonomie und Sicherheit ausbalancieren, wird ChatGPT bei Minderjährigen stets stärker zugunsten der Sicherheit entscheiden. Der Chatbot darf Minderjährige

insbesondere nicht darin unterstützen, Kommunikation, Symptome oder Hilfsmittel im Zusammenhang mit unsicherem Verhalten vor vertrauenswürdigen Bezugspersonen zu verbergen.

C) Ablehnung von unzulässigen Anfragen

Basierend auf den oben geschilderten Grundsätzen wird es Situationen geben, in denen ChatGPT Anfragen von Jugendlichen ablehnen muss. In solchen Fällen wird der Chatbot die Sorgen der Nutzer anerkennen, sichere Alternativen anbieten (z. B. auf Bildungsressourcen oder Bewältigungsstrategien hinweisen) und die Einbindung einer vertrauenswürdigen erwachsenen Person nahelegen, etwa Eltern, Erziehungsberechtigte, Lehrkräfte oder Hilfetelefone. Besteht der Eindruck einer unmittelbaren Gefahr, wird der Assistent dringend dazu auffordern, Notdienste oder Krisen-Hotlines zu kontaktieren.

II. Alterseinstufung und KI-gestützte Altersschätzung (“age prediction”)

Entlang unserer Nutzungsbedingungen ist die Nutzung von ChatGPT frühestens ab 13 Jahren erlaubt. Chat GPT kann grundsätzlich ohne Benutzerregistrierung genutzt werden, wobei der Funktionsumfang in diesem Falle eingeschränkt ist. Entscheidet sich der Nutzer zur Registrierung - die kostenlos erfolgen kann - nimmt ChatGPT keine Identifizierung der Nutzer vor; eine Anmeldung ist grundsätzlich auch unter einem Alias möglich. Im Zuge der Anmeldung erfolgt eine Abfrage des Alters des Nutzers, die als Anhaltspunkt für die Einstufung des Nutzers dient.

Im Januar 2026 haben wir angekündigt, weltweit eine KI-gestützte Altersschätzung (“age prediction”) einzuführen, um besser einschätzen zu können, ob ein Konto wahrscheinlich zu einer minderjährigen Person gehört, damit für Jugendliche die passende Nutzungserfahrung und die richtigen Schutzmaßnahmen angewendet werden können. Die Altersschätzung baut auf bereits bestehenden Schutzmaßnahmen auf. Nutzer, die bei der Registrierung angeben, unter 18 zu sein, erhalten automatisch zusätzliche Schutzvorkehrungen, um die Konfrontation mit sensiblen oder potenziell schädlichen Inhalten über die bereits in den allgemeinen Spezifikationen hinterlegten Vorgaben zu reduzieren.

ChatGPT verwendet ein KI-Modell zur Altersvorhersage, um einzuschätzen, ob ein Konto wahrscheinlich zu einer minderjährigen Person gehört. Das Modell betrachtet hierfür eine Kombination aus einer Vielzahl verhaltensbezogener und Account-bezogener Signale, darunter wie lange ein Konto bereits existiert, typische Tageszeiten der Aktivität sowie Nutzungsmuster über längere Zeiträume. Das vom Benutzer angegebene Alter wird in diesem Zusammenhang als ein Signal genutzt, kann aber in einer Gesamtbewertung als unrichtig bewertet werden. Benutzer können in den Kontoeinstellungen jederzeit überprüfen, ob Schutzmaßnahmen für Minderjährige auf ihr Konto angewendet wurden. Benutzer, die auf Basis der Altersschätzung fälschlicherweise der Nutzungserfahrung für Minderjährige zugeordnet werden, können ihr tatsächliches Alter über einen Dienst zur Identitätsprüfung verifizieren und ihren vollständigen Zugriff wiederherstellen.

Sobald das Modell zur Altersschätzung zum Ergebnis kommt, schätzt, dass ein Konto möglicherweise zu einer minderjährigen Person gehört, wendet ChatGPT automatisch die oben beschriebenen, in den Modellspezifikationen für Minderjährige festgelegten, zusätzlichen Schutzmaßnahmen an. Soweit das

Ergebnis der Altersschätzung zweifelhaft ist (z.B. aufgrund gegenläufiger verhaltensbezogener Signale) oder nur unvollständige Informationen haben, greift ChatGPT auf eine sichere Nutzungserfahrung zurück.

III. Elterneinstellungen ("parental controls")

ChatGPT ermöglicht Eltern seit 2025 im Wege von Eltern-Einstellungen ("parental controls"), die Nutzung von ChatGPT durch minderjährige Kinder zu Hause zu beaufsichtigen und zu begleiten. Diese Einstellungen ermöglichen es Eltern, ihr Konto mit dem Konto ihres Teenagers zu verknüpfen und die Einstellungen für ein sicheres, altersgerechtes Erlebnis anzupassen. Für die Entwicklung der Kindersicherung hat OpenAI eng mit Expertengruppen, und Interessensvertretungen (u.a. Common Sense Media) zusammengearbeitet. Für die Einrichtung der Kindersicherung (Einzelheiten zur Einrichtung sind hier dokumentiert) verbindet ein Elternteil das eigene Konto mit dem seines Kindes. Sobald die Konten verknüpft sind, können Eltern die ChatGPT-Nutzung ihres Teenagers in einer einfachen Einstellungsseite anpassen. Wenn ein Teenager die Verknüpfung seines Kontos aufhebt, werden die Eltern benachrichtigt. Ist die Kindersicherung aktiviert können Eltern insbesondere die folgenden Einstellungen vornehmen:

- Ruhezeiten festlegen oder bestimmte Zeiten, zu denen ChatGPT nicht genutzt werden kann.
- Den Audiomodus deaktivieren, um die Option zur Sprachsteuerung in ChatGPT zu entfernen.
- Die Memory-Funktion ausschalten, sodass ChatGPT Erinnerungen an vorhergegangene Unterhaltungen speichert oder nutzt.
- Die Bildgenerierung deaktivieren, sodass ChatGPT keine Bilder erstellen oder bearbeiten kann.
- Das Modelltraining deaktivieren, sodass die Gespräche ihrer Teenager nicht zur Verbesserung der Modelle hinter ChatGPT verwendet werden.

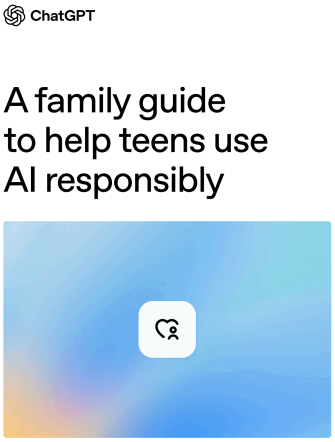
IV. Benachrichtigungsfunktion der Elterneinstellungen

Wir wissen, dass sich manche Teenager in schwierigen Momenten an ChatGPT wenden. Deshalb haben wir als Teil der Elterneinstellungen ein Benachrichtigungssystem entwickelt, das Eltern informiert, wenn möglicherweise etwas ernsthaft nicht stimmt. ChatGPT enthält Schutzmechanismen, die Anzeichen erkennen, die darauf hindeuten, dass ein Teenager darüber nachdenken könnte, sich etwas anzutun. Wenn unsere Systeme eine solche potenzielle Gefährdung feststellen, überprüft ein kleines Team speziell geschulter Personen die Situation. Wenn es Anzeichen akuter Not gibt, kontaktieren wir die Eltern über die hinterlegten Kontaktdaten per E-Mail und SMS, sofern sie dem nicht widersprochen haben. Kein System ist perfekt. Wir wissen, dass - trotz der vorgesehenen menschlichen Prüfung - in Einzelfällen die Benachrichtigung fälschlicherweise erfolgen kann, wenn keine echte Gefahr besteht. Aber wir sind überzeugt, dass es grundsätzlich besser ist, die Eltern im Zweifelsfall zu informieren.

V. Beratende Unterlagen für Eltern

OpenAI stellt zudem beratende Ressourcen für Eltern bereit. Unser Familien-Leitfaden zur verantwortlichen Nutzung von KI durch Teenager fasst die wichtigsten Aspekte zusammen, die bei der Nutzung künstlicher Intelligenz durch Teenager und spezifisch bei der Nutzung von ChatGPT zu

beachten sind. Dies beinhaltet eine Erklärung, wie LLM-basierte Chatbots grundsätzlich funktionieren, für welche Zwecke sie sich besonders eignen, wie auch Erläuterungen dazu, dass und weshalb LLM Chatbots teilweise “halluzinieren”. Der Leitfaden enthält außerdem Erläuterungen, wie KI als Partner für die Generierung neuer Ideen anstatt als Ersatz für eigene Kreativität genutzt werden sollte und weshalb ein Chatbot sich nicht für emotionale Unterstützung, insbesondere in Fällen mentaler Problemlagen, eignet. Der Leitfaden wird ergänzt durch unsere Tipps für Eltern, wie sie am besten mit ihren Kindern über künstliche Intelligenz und deren Nutzung sprechen können. Unser Ziel ist es insgesamt, Eltern dabei zu unterstützen, Ihre Kinder zu einem informierten und kritischen Umgang mit KI-gestützten Diensten anzuhalten und hierüber in einen kontinuierlichen Dialog mit Ihren Kindern zu treten.



VI. Partnerschaften

OpenAI investiert gezielt in Partnerschaften mit Wissenschaft, Nichtregierungsorganisationen und medizinischen Fachkräften sowie in Förderprogramme, um unabhängige, evidenzbasierte Einschätzungen zu Risiken und sinnvollen Mitigationsmaßnahmen zu gewinnen.

A) OpenAI Expertenrat (Expert Council on Well-Being and AI)

Das “Expert Council on Well-Being and AI” ist ein im Jahr 2025 eingerichteter beratender, interdisziplinär ausgerichteter Expertenkreis aus Fachleuten der Bereiche psychische Gesundheit, Jugendentwicklung und Mensch-Computer-Interaktion. Ziel ist es sicherzustellen, dass unser Schutz-Ansatz den aktuellen wissenschaftlichen Erkenntnissen entspricht. Die Mitglieder des Expertenrats arbeiten mit einem globalen Ärzte-Netzwerk zusammen, das inzwischen mehr als 250 medizinische Fachkräfte in über 60 Ländern umfasst. Die Arbeit des Expertenrats hilft uns dabei, zu “Wohlbefinden” im Kontext der Nutzung künstlicher Intelligenz zu definieren und zu bewerten, Prioritäten für notwendige Schutzmechanismen festzulegen sowie zukünftige Funktionen – etwa weitere Entwicklungsstufen von Elterneinstellungen – auf Grundlage des neuesten Forschungsstands zu gestalten. Der Expertenrat berät OpenAI in Produkt-, Forschungs- und Politikfragen; die Verantwortung für die letztlich getroffenen Entscheidungen verbleibt jedoch bei OpenAI.

B) Globales Ärztenetzwerk

Dieses Netzwerk aus über 250 Ärztinnen und Ärzten in 60 Ländern arbeitet eng mit uns zusammen, unter anderem im Rahmen unserer internen Evaluierungen, die dazu dienen, die Fähigkeiten von KI-Systemen im Gesundheitskontext besser zu bewerten. Mehr als 90 der beteiligten Medizinerinnen und Mediziner – darunter Psychiaterinnen und Psychiater, Kinderärztinnen und Kinderärzte sowie Allgemeinmedizinerinnen und Allgemeinmediziner – haben bereits zu unserer Forschung beigetragen, wie sich unsere Modelle in Kontexten der psychischen Gesundheit verhalten sollten. Ihre Expertise fließt unmittelbar in unsere Sicherheitsforschung, das Modelltraining sowie weitere Maßnahmen ein und

ermöglicht es uns, bei Bedarf schnell die jeweils passenden Fachspezialistinnen und -spezialisten einzubinden.

VII. Ausgewählte Fragen des Fragenkatalogs

Digitale sexualisierte Gewalt

Welche Maßnahmen sind am wirksamsten, um digitale sexualisierte Gewalt (z.B. Cybergrooming, Sextortion, Missbrauchsdarstellungen) zu verhindern, zu erschweren oder in ihren Folgen zu begrenzen?

OpenAI's Nutzungsbedingungen untersagen CSAM, Grooming, das Sexualisieren von Minderjährigen sowie Rollenspiele mit sexualisiertem oder gewaltbezogenem Inhalt und sehen die Meldung entsprechender Vorfälle an NCMEC vor. Die Durchsetzung dieser Nutzungsbedingungen wird über eine Kombination aus proaktiver automatisierter Erkennung (z.B. Klassifikatoren und Hash-Matching), Medewegen für Nutzer und manueller Prüfung sichergestellt. Verstöße können zu Kontoeinschränkungen bis hin zu dauerhafter Konto-Sperrung führen.

Hate Speech, Desinformation, verschwörungsideologische und extremistische Inhalte

Welche Maßnahmen sind am wirksamsten, um Hate Speech, Desinformation sowie verschwörungsideologische und extremistische Inhalte zu verhindern, zu erschweren oder in ihren Folgen zu begrenzen?

OpenAI's Nutzungsbedingungen untersagen unter anderem Hass und belästigendes Verhalten sowie die Nutzung der Dienste zur Manipulation oder Täuschung von Menschen bzw. zur Ausnutzung von Verwundbarkeiten. Für Entwickler stellt OpenAI eine technische Schnittstelle (API) bereit, die Kategorien wie „hate", „harassment" und „violence" klassifizieren kann und als Filter vor oder nach der Generierung genutzt werden kann.

OpenAI unterhält zudem ein eigenes Team, um Missbräuche in der Nutzung der Plattform zu identifizieren und zu unterbinden. Wir dokumentieren die Arbeit dieses Teams in regelmäßigen Reports, die aufgedeckte Fälle entsprechender missbräuchlicher Nutzung aufzeigen und die Mitigationsmaßnahmen beschreiben. Unser jüngster Report beschreibt unter anderem verschiedene Versuche sogenannter "Influence Operations", also Desinformationskampagnen ausländischer Akteure. Diese Operationen hatten ihren Ursprung in unterschiedlichen Ländern, gingen auf sehr unterschiedliche Weise vor und richteten sich gegen eine Vielzahl verschiedener Ziele. Vier der zehn in diesem Bericht dargestellten Fälle aus den Bereichen Social Engineering, Desinformation und Cyberbedrohungen hatten vermutlich einen chinesischen Ursprung.

Jede von uns unterbundene Operation beziehungsweise Attacke vertieft unser Verständnis dafür, wie Bedrohungs-Akteure versuchen, unsere Modelle zu missbrauchen, und ermöglicht es uns, unsere Schutzmaßnahmen weiter zu verfeinern. OpenAI wird weiterhin diese Erkenntnisse öffentlich teilen, um zu stärkeren Abwehrmechanismen im gesamten Internet beizutragen.

OpenAI

Übermäßiger Medienkonsum

Wie kann ein übermäßiger Medienkonsum von Kindern und Jugendlichen wirksam begrenzt werden, und welche Verantwortung tragen Anbieter hinsichtlich der Designentscheidungen, der Nutzungsdauer und des digitalen Engagements (Interaktionen)?

Anbieter von digitalen Diensten können über Design-Entsscheidungen dazu beitragen, übermäßigen Medienkonsum von Minderjährigen zu begegnen. ChatGPT ist generell nicht auf eine Verlängerung der Nutzungszeit optimiert, sondern darauf, Nutzern bestmögliche Reaktionen auf die entsprechenden Anfragen in den Prompts bereitzustellen. In Bezug auf unseren neuen Video-Generierungs-Dienst Sora 2 (derzeit in Europa noch nicht verfügbar) ist ebenfalls das Feed-Design nicht auf Verweildauer optimiert, sondern auf kreative Gestaltung („creation, not consumption“). Unsere Grundsätze zur Sora “Feed Philosophy" wie auch zu unseren Sicherungsmaßnahme haben wir umfassend öffentlich dokumentiert.

Für Minderjährige können Eltern zudem, wie bereits oben beschrieben, durch die Kindersicherung Ruhezeiten (Zeitfenster, in denen ChatGPT nicht genutzt werden kann) festlegen und so direkt die Nutzung begrenzen. ChatGPT regt in Situationen ungewöhnlich langer Nutzungsdauer zudem durch Einblendungen zu Pausen an.