# Warum Stable Diffusion *problematische Inhalte* rendern kann – Ein kritisches Handout zur Medienkompetenz

## Einleitung

Stable Diffusion ist ein KI-Modell zur Bildgenerierung, das auf der Verarbeitung großer Mengen an Bild-Text-Paaren basiert. Obwohl das Modell nicht mit illegalen oder expliziten Darstellungen von Kindern trainiert wurde, ist es technisch möglich, dass es durch abstrakte Kombination bekannter Merkmale problematische Inhalte erzeugt. Dieses Handout soll aufklären, warum das möglich ist, was rechtlich relevant ist und warum ein bewusster, reflektierter Umgang mit dieser Technologie notwendig ist.

## Wichtige Klarstellung

**Weder Stable Diffusion noch dessen Trainingsdaten enthalten explizite Darstellungen von Kindern.** Die Trainingsdaten bestehen überwiegend aus öffentlich verfügbaren Bildern, darunter auch normale (nicht-sexualisierte) Darstellungen von Kindern (z.B. aus Stockfotos, Wikipedia, News-Seiten etc.). Explizite Inhalte (z.B. pornografische Bilder) stammen ausschließlich aus Quellen, die Erwachsene zeigen.

## Warum kann das Modell trotzdem problematische Inhalte generieren?

Stable Diffusion arbeitet nicht mit konkretem Bildgedächtnis, sondern mit abstrahierten Repräsentationen von Konzepten (sogenannten *latenten Räumen*). Das bedeutet:

- Das Modell "weiß" z.B., wie der Körperbau, die Kleidung oder die Pose eines Erwachsenen in expliziten Bildern aussieht.
- Es kennt auch normale Darstellungen von Kindern, z.B. Gesichtszüge, Proportionen, Kleidung.
- Wenn ein Prompt fälschlicherweise beide Konzepte kombiniert (z.B. durch missverständliche oder manipulativ formulierte Prompts), kann das Modell eine problematische Darstellung halluzinieren – ohne, dass ein explizites Bild eines Kindes im Training enthalten war.

Diese Fähigkeit zur **Kombination** abstrakter Merkmale ist der Grund, warum KI-Modelle kreative, aber auch ethisch bedenkliche Ergebnisse erzeugen können – ohne "gewollt" dafür programmiert worden zu sein.

## Rechtlicher Hinweis

**Die Erzeugung, Verbreitung oder der Besitz sexualisierter Darstellungen Minderjähriger ist in vielen Ländern (inkl. Österreich) strafbar – auch wenn es sich um KI-generierte Bilder handelt.**

Auch wenn kein reales Kind abgebildet ist, gelten viele solcher Bilder rechtlich als verbotene Darstellung. Der Gesetzgeber betrachtet sie als geeignet, reale Gewalt zu fördern oder zu verharmlosen.

## Zweck dieses Handouts

Dieses Handout ist **kein** Freibrief, sondern eine Warnung. Es soll euch für die Verantwortung beim Umgang mit generativer KI sensibilisieren. Stable Diffusion ist ein mächtiges Werkzeug, das in der Kunst, Wissenschaft und Bildung enorme Chancen bietet – aber nur, wenn es **kritisch reflektiert und ethisch verantwortungsvoll** eingesetzt wird.

## Fazit

- KI-Modelle wie Stable Diffusion können problematische Inhalte erzeugen, obwohl diese nicht im Trainingsmaterial enthalten waren.
- Dies ist technisch erklärbar, aber ethisch und rechtlich hochsensibel.
- Aufklärung, Medienkompetenz und rechtliche Bildung sind zentrale Bausteine, um Missbrauch zu verhindern.