

Deepfakes

Status quo und Handlungsempfehlungen im Strafrecht

Basierend auf der Bachelorarbeit «Erkennen von Deepfakes und KI-generierten Videos in der Strafverfolgung» (Silvan Vifian). Detaillierte Ausführungen und Quellenangaben sind der vollständigen Arbeit zu entnehmen oder beim Autor erhältlich.

Was sind Deepfakes?

Deepfakes sind durch generative künstliche Intelligenz erstellte oder veränderte Medien (Bilder, Videos, Audio), die reale Personen in Situationen darstellen, welche so nie stattgefunden haben. Durch den rasanten technologischen Fortschritt erreichen diese Fälschungen heute ein Qualitätsniveau, das für das blosse menschliche Auge kaum noch als Manipulation zu erkennen ist. Da Gerichte audiovisuellen Aufnahmen traditionell eine hohe Beweiskraft beimessen, wird die zwingend notwendige Unterscheidung zwischen Realität und Fälschung für die Justiz zunehmend anspruchsvoller, was deren Beweiswert künftig gefährdet.

Gefahren

Beweisfälschung

Die Erstellung von Deepfakes im Zusammenhang mit einem Strafverfahren zielt darauf ab, die Wahrheitsfindung durch manipulierte Medien gezielt zu untergraben. Bei Affektdelikten ist dieses Risiko verhältnismässig klein, jedoch können Beweise auch hier im Nachgang manipuliert werden. Ein grösseres Risiko besteht hingegen bei prämeditierten Straftaten, in der organisierten Kriminalität sowie in komplexen Wirtschaftsstrafsachen. In diesen Szenarien nutzen Straftäter die Technologie, um sich mit fingierten Entlastungsbeweisen präventiv ein falsches Alibi zu verschaffen, oder sie generieren kompromittierendes Material, um unschuldige Dritte gezielt mit Vorwurfsbeweisen zu belasten.

Liar's dividend

Die blosse Existenz hochauflösender Manipulationswerkzeuge etabliert eine neue Verteidigungsstrategie, die in der Literatur als «*Liar's Dividend*» bezeichnet wird. Dieses Phänomen beschreibt die gezielte Taktik, die Authentizität von echtem, belastendem Beweismaterial systematisch als KI-generiert infrage zu stellen. Dass diese Strategie bereits heute angewandt wird, illustriert beispielsweise der Fall aus den USA *Huang v. Tesla Inc. (2023)*. Solche pauschalen Schutzbehauptungen erzwingen aufwändige forensische Abklärungen. Dies führt unweigerlich zu Verfahrensverzögerungen und droht, die personellen Ressourcen der digitalen Forensik künftig massiv zu überfordern.

Deepfakes können auch eigenständige Straftatbestände erfüllen (z.B. Rachepornografie). Diese Thematik wird in diesem Factsheet nicht behandelt, obschon sich in der Praxis ähnliche Herausforderungen stellen.

Status quo Justiz

In der internationalen Rechtsprechung finden sich erst wenige dokumentierte Präzedenzfälle zu Deepfakes als Beweismittel, wie etwa die Verfahren *Mendones v. Cushman & Wakefield, Inc. (2025)* und *Matter of M.S. (2026)*. Im europäischen und im Schweizer Rechtsraum sind bislang keine vergleichbaren Fälle bekannt. Dementsprechend existieren derzeit weder spezifische gesetzliche Regelungen noch behördliche Leitfäden für den verfahrensrechtlichen Umgang mit synthetischen Medien.

Der Diskurs mit Schweizer Fachexperten zeigt, dass das Prinzip der freien Beweiswürdigung praxisintern weiterhin als technologieneutral und dogmatisch ausreichend erachtet wird. Es wird darauf vertraut, dass isolierte Fälschungen auffallen, da sie unweigerlich in Widerspruch zur Gesamtgeschichte und der übrigen Spurenlage geraten. Ob diese klassische Plausibilitätsprüfung künftig auch bei hochprofessionellen Fälschungen als stärkster Filter genügt, bleibt jedoch kritisch zu beobachten. In der Praxis erfolgt bei Verdacht eine Delegation an die digitale Forensik. Als grösste prozessuale Gefahr gilt übereinstimmend die taktische Schutzbehauptung (*Liar's Dividend*).

Status quo
Generierung

Die technologischen Fähigkeiten zur Erstellung synthetischer Medien haben sich in den letzten Jahren rasant weiterentwickelt, primär getrieben durch neue Architekturen wie Diffusionsmodelle und Transformer. Zwar stossen aktuelle Manipulationspipelines bei qualitativ minderwertigem oder verdecktem Überwachungsmaterial (CCTV) teilweise noch an ihre Grenzen. Es ist jedoch lediglich eine Frage der Zeit, bis diese Hürden überwunden werden, da die methodischen Fortschritte aus der statischen Bildgenerierung rasch auf die Videobearbeitung konvergieren. Dass automatisierte und überzeugende Identitätseingriffe unter echten, stark komprimierten Bedingungen bereits heute erschreckend praxistauglich sind, verdeutlicht der direkte Vergleich der untenstehenden Bildausschnitte.



Originales Video



Deepfake Video

Status quo
Erkennung &
Forensik

Obwohl die Forschung stetig neue Detektionsmodelle hervorbringt, fehlt es diesen massiv an Generalisierungsfähigkeit. Die Systeme sind zu stark auf spezifische, ihnen bereits bekannte Fälschungstypen optimiert. Die Tabelle verdeutlicht dies anhand der AUC-Skala: Ein Wert von 1.0 bedeutet, jedes Video perfekt zu erkennen, während 0.50 genauso gut ist wie zufälliges Raten. Während die Modelle bei vertrauten Generierungsmethoden exzellent abschneiden, bricht ihre Leistung bei einem abweichenden oder neuartigen Fälschungstypus drastisch in Richtung dieses Zufallswerts ein. Für die Beweisführung vor Gericht bieten rein algorithmische Detektoren aufgrund dieser Unzuverlässigkeit derzeit keine hinreichende Sicherheit.

Die forensische Praxis stützt sich bei Verdachtsfällen daher vorerst primär auf bewährte, IT-Forensik wie Metadaten- und Herkunftsanalysen. Für tiefergehende, KI-spezifische Bildanalysen bedarf es jedoch hochspezialisierter Fachpersonen, von denen in der Branche aktuell erst eine überschaubare Anzahl zur Verfügung steht. Angesichts der rasanten technologischen Evolution der Deepfakes birgt dies ein potenzielles systemisches Risiko. Insbesondere durch taktische Schutzbehauptungen droht eine Zunahme an Prüfaufträgen. Für die Justiz bedeutet dies einen massiven Mehrbedarf an personellen und technischen Ressourcen, zumal die Erstellung spezialisierter forensischer Gutachten sehr hohe Kosten verursacht.

Modell	Originale Daten (AUC)	Abweichende Daten (AUC)	Differenz (Δ)
Action-Guided-Units	97.15%	91.00%	-6.15%
FakeSTormer	98.5%	74.20%	-24.3%
LAA-Net	96.30%	55.20%	-41.10%

Handlungsempfehlungen

1
Problembewusstsein
Die Justiz und Anwaltschaft müssen das Risiko manipulierter Beweismittel proaktiv adressieren und Verdachtsmomente konsequent äussern. Ein solcher Manipulationsverdacht ist stets kritisch zu prüfen, wobei pauschale Deepfake-Vorwürfe auch als rein verfahrenstaktische Schutzbehauptungen (*Liar's Dividend*) zu würdigen sind.

2
Ausbau der digitalen Forensik
IT-forensische Stellen müssen ihre Kompetenzen zur Deepfake-Erkennung proaktiv ausbauen, anstatt primär reaktiv zu agieren. Angesichts der rasanten technologischen Entwicklung ist dies zwingend erforderlich, um künftige Prüfaufträge rechtzeitig und verlässlich bewältigen zu können.

3
Richtlinien & Standards
Anstelle von Gesetzesrevisionen bedarf die Justiz zwingend einheitlicher Praxisrichtlinien und behördeninterner Weisungen. Solche standardisierten Leitfäden definieren den konkreten Umgang mit KI-Manipulationsvorwürfen und sichern so eine rechtsgleiche und personenneutrale Prozessführung.