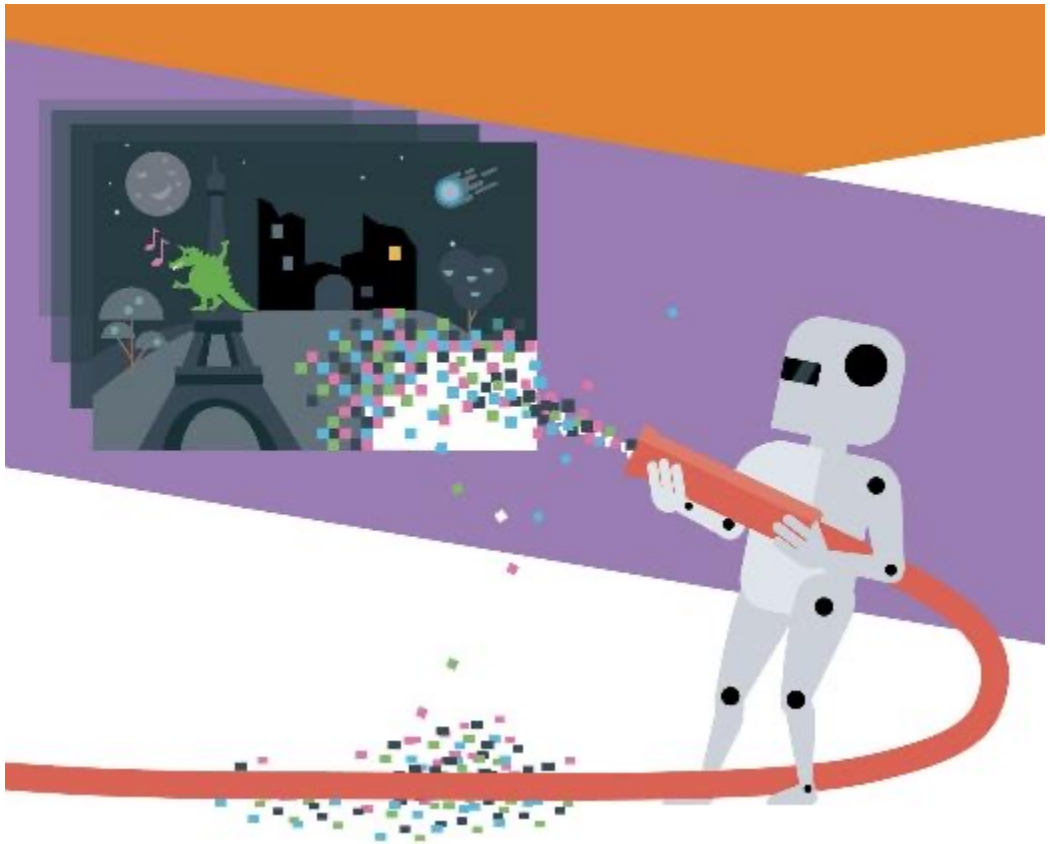
Smart City Verein Bern «AI for Purpose»

Deepfakes: Handlungsempfehlungen für die Schweizer Gesellschaft



Laetitia Ramelet

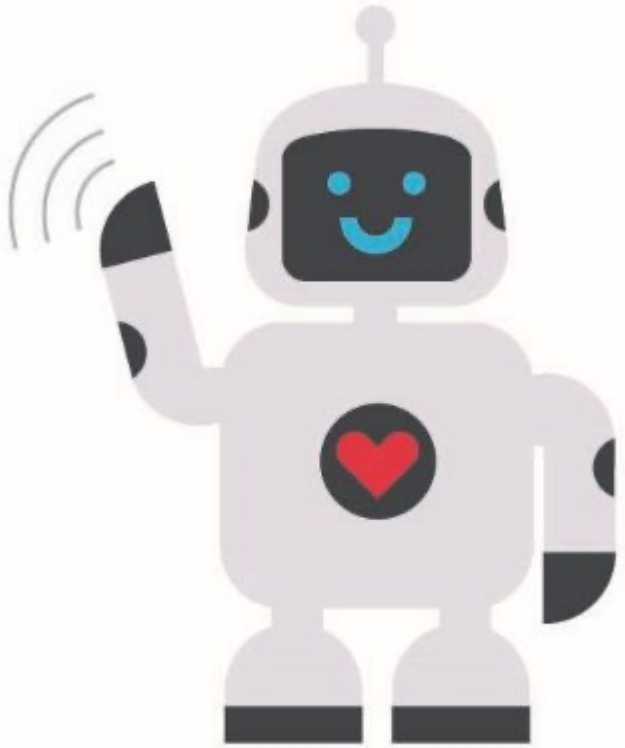
TA-SWISS

mitglied der akademien der wissenschaften schweiz

Illustrationen: Hannes Saxer, Bern

TA-SWISS
Stiftung für Technologiefolgen-Abschätzung

- **Abschätzung der Chancen und Risiken neuer Technologien für die Gesellschaft**
- **Sachliche und ausgewogene Studien** für Parlament, Bundesrat, Verwaltung und Stimmbevölkerung
- **Interdisziplinäre Perspektive**
- **Auftrag im Bundesgesetz über die Förderung der Forschung und der Innovation (Art. 11)**



«Deepfakes und manipulierte Realitäten»

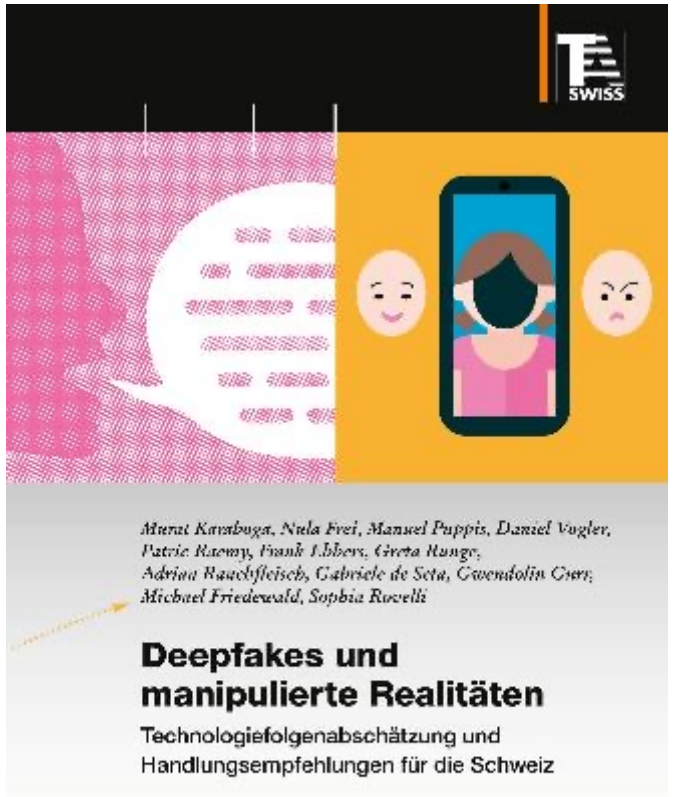
Studie im Auftrag von TA-SWISS (2024)

«**Deepfakes und manipulierte Realitäten.**
Technologiefolgenabschätzung und
Handlungsempfehlungen für die Schweiz»

AutorInnen: Murat Karaboga, Nula Frei, Manuel Puppis, Daniel Vogler, Patric Raemy, Frank Ebbers, Greta Runge, Adrian Rauchfleisch, Gabriele de Seta, Gwendolyn Gurr, Michael Friedewald, Sophia Rovelli

Fraunhofer- Institut für System- und Innovationsforschung ISI (Karlsruhe), Universität Freiburg i.Ue., Universität Zürich

Zur Studie: https://www.ta-swiss.ch/deepfakes



«Ein Deepfake ist ein **mithilfe von KI-Techniken synthetisierter oder manipulierter Audio-, Bild- bzw. Videoinhalt**, der authentisch wirkt, es aber nicht ist.»

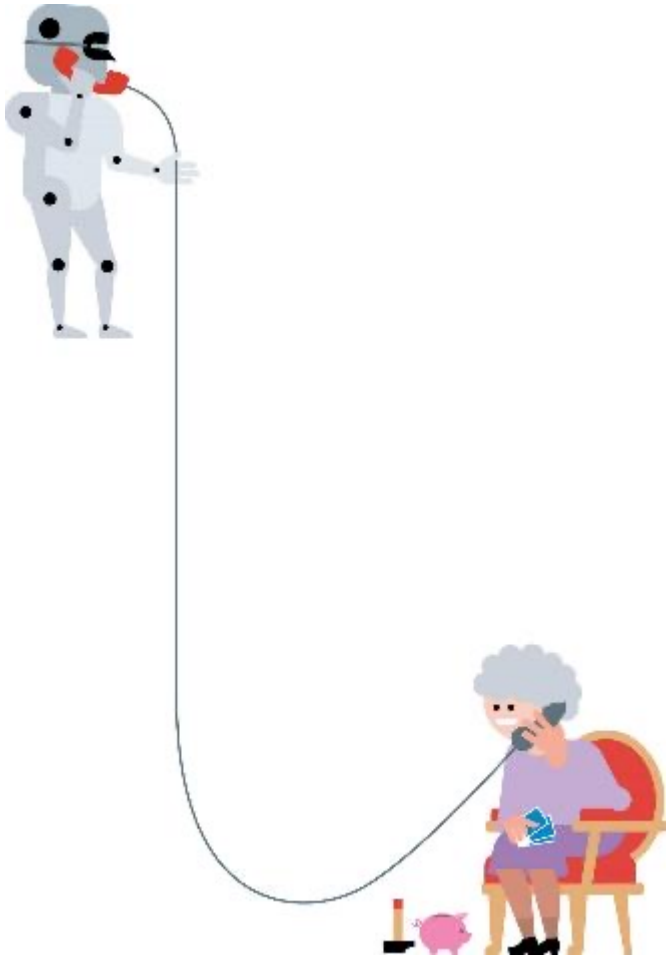
Oft wird ein Mensch gezeigt, «**der etwas sagt oder tut, was er nie gesagt oder getan hat.»**

(Karaboga et al., TA-SWISS)

Quellen: https://www.instagram.com/julian_ai_art/; X; Blick.ch.

Quellen:.tokyo; https://thispersondoesnotexist.com/, EPFL Pavillons.

Risiken und Missbräuche…



- Pornographische Deepfakes und Cybermobbing
- Identitätsdiebstahl
- Schockanrufe
- Verbreitung von falschen Informationen und Manipulation
- Cyberattacken

… und Chancen

- Unterhaltungsindustrie
- Werbung und Kommunikation
- Innovative Formate für den Unterricht
- Schutz von Identitäten

Stand der Technik

- Deepfake-Bilder und -Audios relativ leicht zu erstellen
- Videos: (derzeit) viel aufwändiger
- Text-to-Video?

Wie gut können wir Deepfakes von realen Videos unterscheiden?

- Online-Experiment im September 2023 mit 1361 Personen in der Schweiz
- Davon die Hälfte mit einer *Literacy-Intervention* (Tipps), die andere Hälfte ohne
- «Gut gemachte » Deepfakes von bekannten Persönlichkeiten

-> Unterscheidung schon jetzt sehr schwierig

-> auch mit *Literacy-Intervention*

-> positiver Einfluss von Social-Media-Kompetenzen

Aktuelle technische Lösungen

Kein technisches Wundermittel…

- Authentifizierung von originalen Inhalten (mit digitaler Signatur)
- Kennzeichnung von Deepfakes (durch Urheber und Verbreitungskanäle)
- Deepfake-Detektoren

Ein neues Tool für die Cyberkriminalität

Cyberrisiken für Unternehmen und Verwaltungsstellen

- Überwindung von Sicherheitsmassnahmen (insb. mit biometrischen Verfahren)
- Social-Engineering-Angriffe
- Rufschädigung
- Ausserdem:

- Falsche Profile auf sozialen Medien
- Digitales Astroturfing

Schutzmassnahmen für Unternehmen und Verwaltungsstellen

- Sensibilisierung von Mitarbeitenden
- Risikoanalyse innerhalb der Organisation
- Notfall- und Krisenplan
- Vorsicht mit Gesichts- oder Sprachbiometrie
- Freiwillige Meldung von Deepfake-Vorfällen

Allerdings: schneller Wandel und kein Wundermittel

Einige Empfehlungen der TA-Studie

- **Onlineplattformen regulieren**:

Zusammenarbeit mit den Strafverfolgungsbehörden, Meldesystem, Sperrung gemeldeter Deepfakes (bei Verdacht), Transparenz und Widerspruchsmöglichkeiten

- **Bildung und Selbstverantwortung der Bürgerinnen und Bürger fördern**:

Medienkompetenzen und Sensibilisierung beim Teilen von Daten und Deepfakes

- **Beratungszentren für Opfer von Cyberkriminalität unterstützen**

- **Schweizer Unternehmen und Institutionen sensibilisieren und Präventionsmassnahmen ergreifen**:

-> interne Beurteilung der Risiken, Weiterbildung, Plan für den Umgang mit allfälligen schädlichen Deepfakes, fortschrittliche Authentifizierungsmassnahmen

- **Journalistische Standards hochhalten**