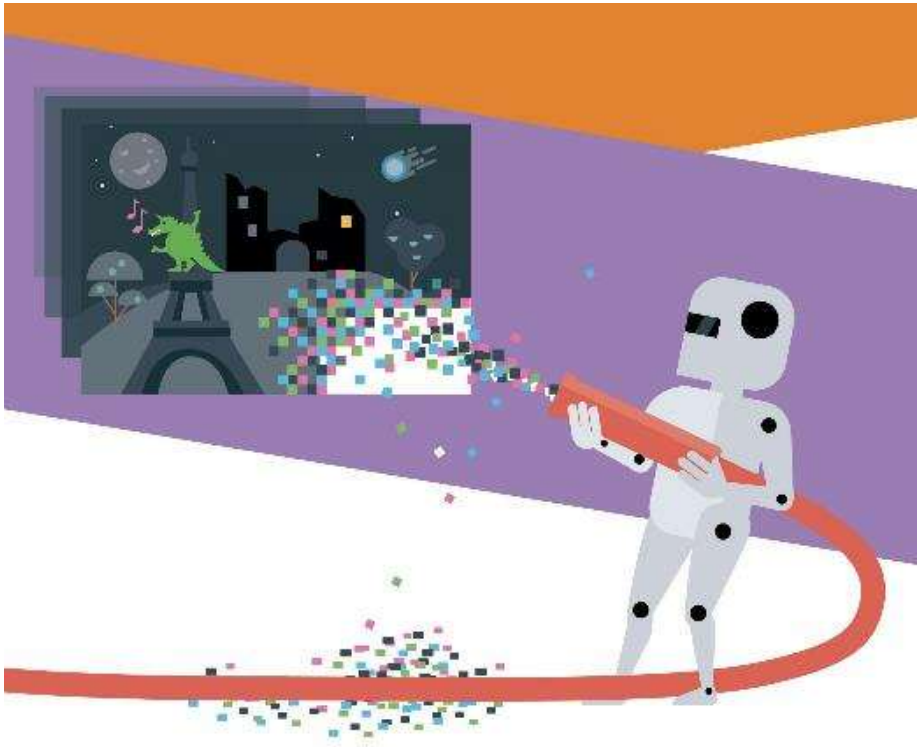
Deepfakes et esprit critique à l'heure de l'IA

Deepfakes et réalités manipulées : résultats de l'étude de TA-SWISS

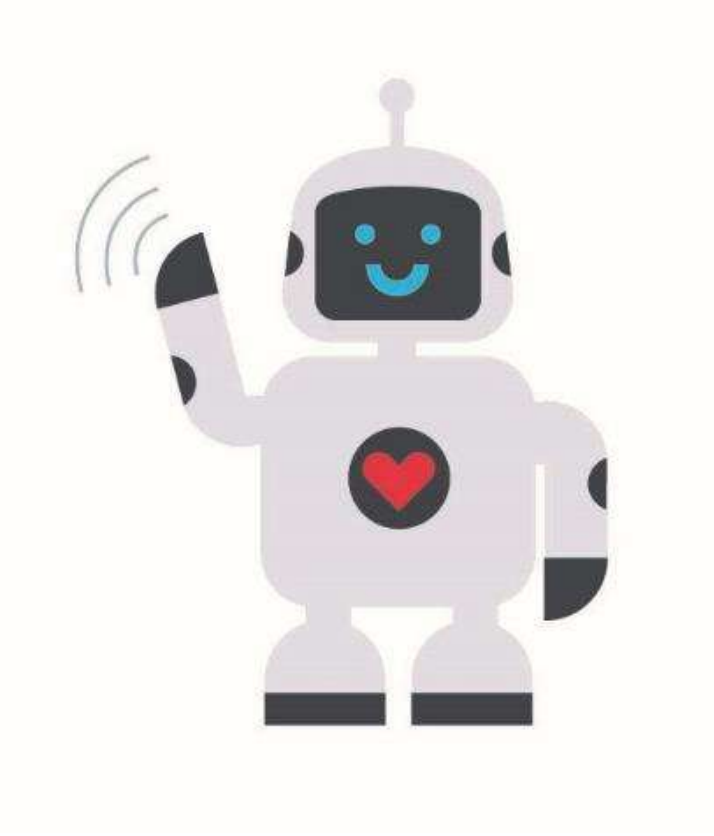


Laetitia Ramelet
TA-SWISS

mitglied der
akademien der wissenschaften schweiz

La Fondation TA-SWISS

- **Évaluation des choix technologiques** (***technology assessment***, TA)
- **Etudes scientifiques sur les opportunités et risques de nouvelles technologies**
- **Approche interdisciplinaire et participative**
- **Mandat ancré dans la loi fédérale sur l'encouragement de la recherche et de l'innovation** (art. 11)



La Fondation TA-SWISS

Mission :

- **Fournir des analyses scientifiques et indépendantes** au Parlement, au Conseil fédéral, aux administrations, à la société civile, l'économie et l'ensemble de la population en Suisse
- **Impliquer les citoyennes et citoyens dans le débat technologique**
- **Prendre en compte les différents points de vue** des parties concernées
- **Diffuser les résultats** à la politique et l'ensemble de la société

« Deepfakes et réalités manipulées » (2024)

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

Auteur(e)s:
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

Télécharger l'étude et sa synthèse: https://www.ta-swiss.ch/fr/deepfakes

«Un **deepfake**, ou hypertrucage, est un contenu **audio ou visuel** (animé) synthétique ou manipulé à l'aide de techniques d'IA,

qui semble **authentique**

et qui, la plupart du temps, fait faire ou dire à un individu quelque chose qu'il **n'a jamais fait ou dit**. **»**

(Karaboga et al., TA-SWISS)

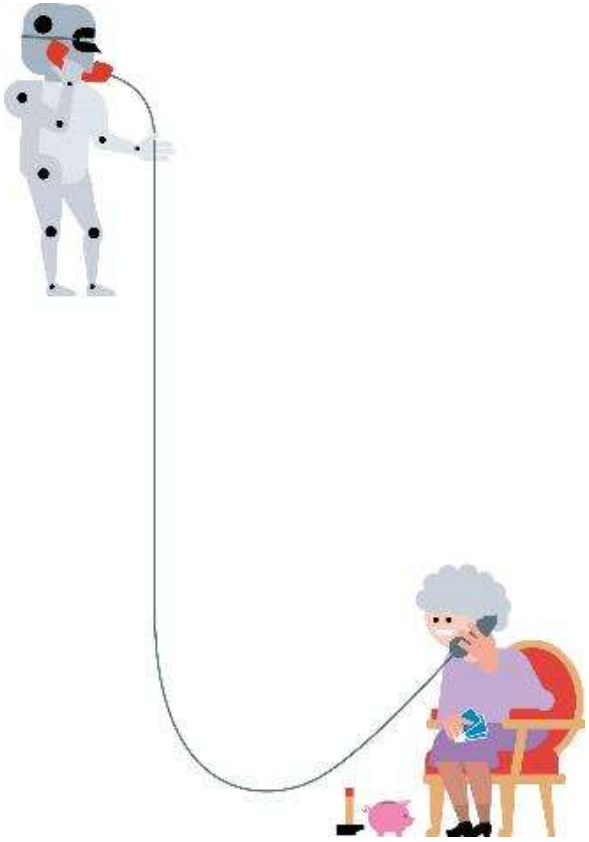
Sources : https://www.instagram.com/julian_ai_art/; X; Blick.ch.

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

Deepfakes: images et vidéos

- ***facial reenactment*** : manipulation des expressions faciales
- ***face morphing*** : fusion de plusieurs visages
- ***face swapping*** : remplacement d'un visage par un autre
- ***face generation*** : création de visages qui n'existent pas dans la réalité
- ***full body puppetry*** : modification des mouvements du corps
- **Générateurs**

Risques et abus…



- Deepfakes pornographiques et cybermobbing
- Vol d'identité
- « Appels chocs »
- Diffusion d'informations fausses et manipulations
- Cyberattaques

… et opportunités

- Industrie du divertissement
- Publicité et communication
- Formats innovateurs pour l'enseignement
- Protection d’identité

Techniques actuelles

- Images et audios relativement faciles à produire
- Vidéos: plus difficiles (pour le moment)
- *Text-to-video*?

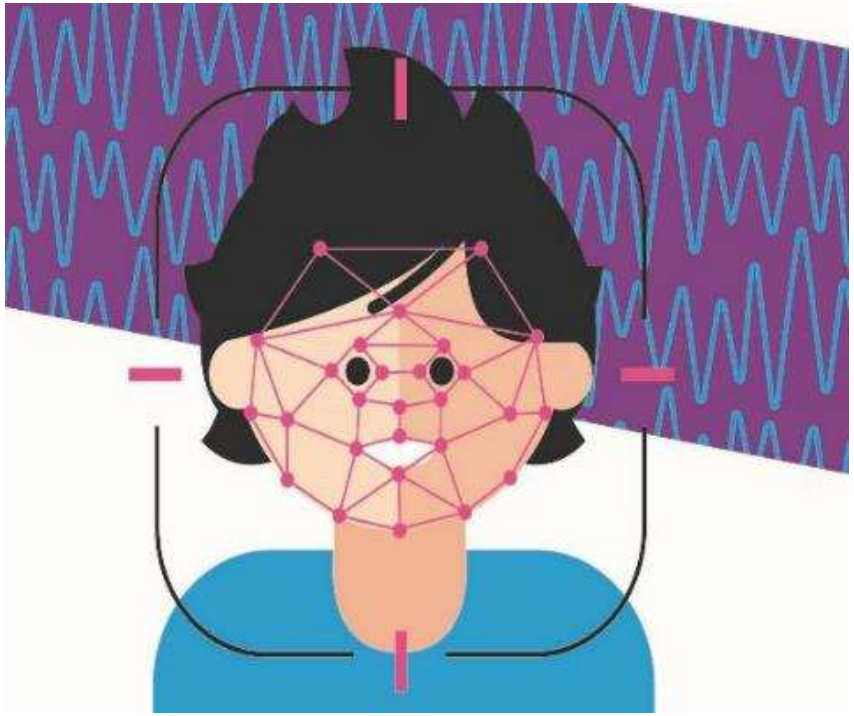
Arrivons-nous encore à voir la différence ?

- Expérience en ligne en septembre 2023 avec 1361 personnes en Suisse
- Astuces données à la moitié (*Literacy-Intervention*)
- Deepfakes de personnalités connues, « bien faits »

-> Distinction déjà très difficile

-> y compris avec *Literacy-Intervention*

-> influence positive des compétences en matière de réseaux sociaux



Solutions techniques actuelles

Pas de remède technique miracle…

- Authentification des contenus originaux (signature numérique)
- Marquage des contenus deepfakes
- Détecteurs de deepfakes

Risques des deepfakes en politique

- Deepfakes de personnalités et partis politiques
- Diffusion de fausses informations
- Outil de chantage et d'intimidation
- Incitation à la violence ou attisement de tensions sociales
- Faux profils sur les réseaux sociaux
- *Astroturfing* numérique
- Dividende du mensonge
- Perte de foi dans les institutions démocratiques ou les médias

Opportunités des deepfakes en politique

- Humour, divertissement et satire en tant que moyens de délibération

- Nouveaux formats pour l'éducation à la citoyenneté ou la participation politique

Les deepfakes et les médias en Suisse

- Un défi pour l'information
- Une opportunité pour le journalisme

- Devoir d'identification rapide
- Devoir de contextualisation
- Quand faut-il parler d’un deepfake ?

- Outil d'intimidation

Entretiens avec des institutions de formation au journalisme et des rédactions (2022)

- **Valeur des normes de base du journalisme**

➔ **à soigner en pratique et dans la formation**

- Deepfakes souvent traités comme un cas de désinformation
- Besoin de spécialistes pour les cas techniquement complexes ? (vs. ressources limitées)
- Crainte d'une perte de confiance dans les médias
- Perception d'une opportunité pour les médias

➔ **Besoin d'une sensibilisation du public à la manipulation et à l'importance de la vérification des sources**

Quelques recommandations de l'étude TA

- **Réglementation des plateformes en ligne:**

Coopération avec les autorités de poursuites pénales, système de notification, blocage des deepfakes illicites, directives de transparence et mécanismes de recours

- **Responsabilité personnelle et formation des citoyennes et citoyens**:

promotion des compétences médiatiques, sensibilisation aux enjeux du partage en ligne
→ esprit critique

- **Soutien aux centres de consultation pour les victimes de cybercriminalité**

- **Mesures de prévention dans les organisations suisses:**

Evaluation des risques, **formation continue,** mécanismes de gestion d'un éventuel deepfake, systèmes d'authentification avancés

- **Maintien de normes journalistiques élevées**

Deepfakes et esprit critique ?

- Une période de transition
- Pas d'apocalypse de l'information pour l'instant
- Redéfinition des frontières entre réel et virtuel?
- Combien de véracité et d'authenticité voulons-nous ?
- Comment voulons-nous utiliser l'IA pour représenter le monde ?