Brunngasse 36
CH-3011 Bern
www.ta-swiss.ch

Documents concernant la mise au concours de l'étude « Deepfakes et réalités manipulées »

Délai pour le dépôt des esquisses de projet : **16 avril 2021**

Délai pour le dépôt des dossiers de candidature : **22 août 2021**

mitglied der akademien der wissenschaften schweiz

Stiftung für Technologiefolgen-Abschätzung
Fondation pour l'évaluation des choix technologiques
Fondazione per la valutazione delle scelte tecnologiche
Foundation for Technology Assessment

1. Themenbeschreibung: Deepfakes und manipulierte Realitäten: Wenn wir unseren Augen nicht trauen können

Etwas mit eigenen Augen gesehen zu haben, galt – zumindest bei Menschen, deren Wahrnehmung nicht durch Drogen oder Wahnvorstellungen getrübt ist – als Prüfstein für Realität und Wahrheit schlechthin. Im Zeitalter von geschönten Realitäten, manipulierten Fotos und Videos stellt sich jedoch die Frage, ob wir auch künftig noch unseren Augen trauen können. Dank Bildbearbeitungssoftware und künstlicher Intelligenz lassen sich heute Ton, Bilder und Videos auf eine nahezu unsichtbare Weise manipulieren. Ob die Realität wiedergegeben wird oder etwas Konstruiertes, lässt sich kaum mehr erkennen.

In der Werbung und in Modemagazinen sind retuschierte Fotos bereits Alltag, inzwischen aber können auch Laien Bilder manipulieren. Täglich laden Nutzer und Nutzerinnen auf Instagram 95 Millionen Fotos und Videos hoch und setzen sich dabei gerne ins beste Licht. Dank farbverändernden Filtern und Bearbeitungstools wie Photoshop oder FaceTune lässt sich heute einfach eine optimierte Scheinwelt schaffen. Ein paar Klicks sorgen für eine schmale Taille und einen makellosen Teint. Die Retuschen sind für Laien kaum erkennbar. Auf Instagram scheinen alle glücklich und erfolgreich – und dank Software auch besonders attraktiv.

Künstliche und gefilterte Gesichter beeinflussen auch unsere Vorstellung von Schönheit. Junge Frauen auf der ganzen Welt wollen Gesichtszüge, wie man sie von den geschönten Influencer-Fotos kennt: Winzige Nasen, ausgeprägte Wangenknochen und volle Lippen. Der Psychologe Helmut Leder geht davon aus, dass sich das Schönheitsideal derzeit so schnell verändert wie nie zuvor[1].

Studien zeigen: Wer sich ständig mit geschönten Bildern vergleicht, betrachtet sich selbst kritischer. Soziale Medien können daher depressiv, einsam und unglücklich machen – betroffen sind vor allem Jugendliche. Für manche ist es vom Filter zur Schönheits-OP nur ein kleiner Schritt. Als «Snapchat Dysmorphia» beschreibt eine Untersuchung des Boston Medical Center das Phänomen, dass Menschen zum Chirurgen gehen, um so schön auszusehen wie ihre gefilterten Selbstporträts[2].

Doch nicht nur Bilder, auch Videos können inzwischen so manipuliert werden, dass sie eine veränderte Realität täuschend echt wiedergeben können. **Deepfakes** (ein Kofferwort aus den Begriffen «Deep Learning» und «Fake») beschreiben realistisch wirkende Medieninhalte (Foto, Audio und Video), welche durch Techniken der künstlichen Intelligenz (KI) generiert, abgeändert

[1] https://psychologie.univie.ac.at/news-medienbeitraege/medienbeitraege/details/news/instagram-face-als-realitaet/?tx_news_pi1%5Bcontroller%5D=News&tx_news_pi1%5Baction%5D=detail&cHash=f42a922b34c0ddbb2e2c3890b6d1072c

[2] https://www.sciencedaily.com/releases/2018/08/180802141601.htm

oder verfälscht worden sind. Mit viel Rechenaufwand und angesammelten Datenmengen erschafft die Maschine ein künstliches Video und kreiert eine täuschend echte Kopie der Stimme der betroffenen Person.

Ein neues Phänomen?

Bildmanipulationen sind beinahe so alt wie die Fotografie selbst. Eines der bekanntesten Porträts von Abraham Lincoln beispielsweise war eine Neu-Komposition, wobei sein Kopf auf den Körper eines anderen Staatsmannes (ausgerechnet eines Sklaverei-Befürworters) gesetzt worden war. Autokratische Machthaber wie Lenin, Stalin, Mao Tse-tung oder Hitler liessen unliebsame Personen aus offiziellen Fotos herausretuschieren und damit letztlich aus der Geschichte entfernen. In anderen Fällen wurden Personen hinzugefügt: Während des Präsidentschafts-wahlkampfes 2004 wurde ein Bild veröffentlicht, das John Kerry und Jane Fonda zeigte, wie sie gemeinsam bei einer Anti-Vietnamkrieg-Demonstration sprachen. Das Bild entpuppte sich als politisch motivierte Fälschung.

Auch Nachrichtenbilder wurden schon in der Vergangenheit manipuliert. Nach dem Attentat auf Touristen in Luxor (1997) wurde beispielsweise auf einem Pressefoto eine Wasserlache rot eingefärbt, um blutig zu erscheinen. Im Jahr 2006 veröffentlichte die Nachrichtenagentur Reuters das Foto einer libanesischen Stadt nach einem Bombenangriff, dem mehr Rauch hinzugefügt worden war.

Bis vor kurzem erforderte die Erstellung von überzeugenden Fälschungen viel Zeit und Geschick. Heutzutage ist die Fotobearbeitungssoftware so ausgereift, dass fast jeder mit Zugang zu einem Computer eine überzeugende Fotofälschung erstellen kann. Dank künstlicher neuronaler Netzwerke können solche Fälschungen nun auch weitgehend autonom erzeugt werden. So kann eine Software inzwischen ein computergeneriertes fotorealistisches Gesicht erzeugen, das menschlich aussieht, aber nicht wirklich echt ist[3].

Die rasante Entwicklung wurde durch sogenannte Generative Adversarial Networks (GAN) ermöglicht. Hierbei werden zwei neuronale Netze miteinander kombiniert, von denen das eine System versucht, etwas zu erzeugen (z. B. fiktive Bilder), was durch das andere System bewertet wird (beispielsweise wird also versucht zu bestimmen, ob es sich um ein echtes Bild oder eine Fälschung handelt). Durch die Rückmeldung der Bewertung (und den Lerneffekt) erzeugt das gestaltende Netzwerk immer bessere Vorschläge, sodass das Training schliesslich zu einem täuschend echten Ergebnis führt.

[3] Beispielsweise auf der Webseite thispersondoesnotexist.com.

Nicht mehr nur grosse Filmstudios erschaffen «Realität»

Während Amateure bereits seit einigen Jahren Fotos unkompliziert bearbeiten und retuschieren können, war die fotorealistische Manipulation von Bewegtbildern aufgrund des grossen Aufwands bisher professionellen (Film-)Studios und Spezialisten für Visual Effects vorbehalten. So bildete beispielsweise «Lucasfilm» für die Star Wars-Serie die junge Prinzessin Leia aus der ersten Folge von 1999 digital nach und integrierte sie in die Folge «Episode IV» von 2015. Ein Jahr später hatte in «Rogue One» der bereits 1994 verstorbene Schauspieler Peter Cushing einen Auftritt. Für diese cineastischen Kunstkniffe kam CGI zum Einsatz. Das Kürzel steht für Computer Generated Imagery, ein äusserst aufwendiges Verfahren. Dank der modernen Technologie sollen nun vermehrt verstorbene Personen posthum Rollen in Spielfilmen übernehmen, wie beispielsweise James Dean.

Die Verfügbarkeit von Deep-Learning-Methoden hat nun jedoch auch die Welt der Bewegtbilder demokratisiert: Die Herstellung von Deepfake-Videos ist exponentiell schneller, einfacher und billiger geworden – dadurch können heute auch «Normalsterbliche» eigene Filme erschaffen.

Eine beliebte Spielart von Deepfakes sind Face Swaps: Das gesamte Gesicht einer Person wird durch ein anderes ersetzt, dessen Originalmimik aber beibehalten. Als Grundlage dienen neben dem Originalvideo möglichst viele Bilder der Zielperson, und zwar idealerweise aus unterschiedlichen Perspektiven und mit verschiedenen Gesichtsausdrücken. So können Programme wie FakeApp die wesentlichen Merkmale der beiden Gesichter zunächst erkennen und in der Folge gegeneinander austauschen. Die Palette möglicher Anwendungen reicht dabei von lustigen Spielereien bis zu Missbrauch. Während die einen Schauspieler in berühmten Filmszenen austauschen, weil sie etwa immer schon Sylvester Stallone für den besseren Terminator gehalten haben, verletzen die anderen auf diese Weise Persönlichkeitsrechte, indem sie Gesichter von beliebigen Frauen auf die Körper von Pornodarstellerinnen setzen.

Aufgrund der mittlerweile frei verfügbaren Softwaretools und des technischen Fortschritts (mit steigender Rechenleistung) ist zu erwarten, dass in absehbarer Zeit vermehrt hochauflösende, fotorealistische synthetische Medien und Deepfakes in Umlauf kommen. Angesichts der raschen Entwicklung gehen Fachleute zudem davon aus, dass Deepfakes in wenigen Jahren so perfekt sein werden, dass die Manipulation praktisch nicht mehr nachweisbar sein wird. Die ethischen und gesellschaftlichen Folgen dieser technischen Möglichkeit sind beträchtlich.

Herausforderung und Chance für Strafverfolgung

Deepfakes bieten Cyberkriminellen das Potenzial für umfangreichen Betrug. Gefälschte Audiodateien oder Videos sind ein ideales Werkzeug, um in Phishing-Kampagnen an vertrauliche Informationen heranzukommen. Wenn sich zum Beispiel jemand via Video und mit Einsatz von Deepfake-Technologie überzeugend als Vorgesetzter ausgeben und so nach Passwörtern oder sensiblen Daten fragen kann, oder wenn ein gefälschter Anruf der Chefin mit realistischer

Stimme eine Überweisung in Auftrag gibt. Auch Sicherheitssysteme, die auf Videoentsperrungen beruhen, könnten durch Deepfakes getäuscht werden.

KI-Verfahren könnten den Sicherheitsbehörden andererseits aber auch als Instrument bei der Strafverfolgung, Ermittlung und Analyse helfen. So soll die deutsche Polizei im Kampf gegen sexuellen Kindesmissbrauch computergenerierte kinderpornografische Bilder erstellen dürfen, um effektiver gegen verdächtige Anbieter vorgehen zu können. Denn Zutritt zu solchen Foren erhalten User oft erst, wenn sie selbst Bilder oder Videos hochladen.

Im realen Politbetrieb angekommen

Mittlerweile sind gefälschte Videos auch auf der politischen Bühne in Erscheinung getreten. 2018 sorgte ein Deepfake-Video für Furore, worin der ehemalige US-amerikanische Präsident Barack Obama Donald Trump als Vollidioten bezeichnete. Hinter dem Video steckte der Regisseur und Schauspieler Jordan Peele. Er wollte damit auf die Gefahren manipulierter Videos hinweisen. Und er zeigte gleichzeitig, wie Politikerinnen und Politiker für Deepfakes missbraucht werden können.

Letztes Jahr kursierte auf Facebook ein Video der demokratischen US-Politikerin Nancy Pelosi bei einer Rede. Dabei wurde die Abspielgeschwindigkeit um etwa 75 Prozent verlangsamt; dadurch wirkte die Politikerin betrunken, oder als stünde sie unter Medikamenteneinfluss.

Ein indischer Abgeordneter produzierte sogar Deepfakes von sich selbst, in denen er potenzielle Wähler und Wählerinnen in verschiedenen Sprachen ansprach (die er eigentlich nicht beherrscht).

Angesichts dieser Beispiele scheint es nur eine Frage der Zeit, bis solche Technologien auch zur Diffamierung politischer Gegner eingesetzt werden. Eine falsche Botschaft, die scheinbar aus dem Mund eines Kontrahenten stammt, könnte beispielsweise wahlentscheidend sein oder internationale Konflikte auslösen.

Der Deutsche Bundestag erachtet daher Deepfakes als eine grosse Gefahr für Gesellschaft und Politik, wenn sie dazu genutzt werden, die öffentliche Meinung zu manipulieren und den politischen Prozess gezielt zu beeinflussen. Es könne nicht ausgeschlossen werden, dass aufgrund der schnellen technologischen Entwicklungen künftig auch eine Bedrohung demokratischer Prozesse von Deepfakes ausgehen kann.

Cybermobbing gegen Frauen und Minderheiten

Neben der medial vieldiskutierten Gefahr, die Deepfakes für die Politik darstellen, warnen Technikethiker und Menschenrechtsaktivisten mittlerweile auch vor einem weiteren Problem: Die KI-Verfahren könnten dazu verwendet werden, auch Bürgerinnen und Bürger zu verleumden und zu attackieren.

Ein Beispiel dafür ist eine inzwischen eingestellte App namens «DeepNude». Diese ermöglichte es, Bilder von Frauen zu verwenden, um diese virtuell «auszuziehen». Die Software tauscht die Kleidung von Frauen auf einem Foto durch realistische nackte Körper aus. Obwohl die Deepfakes keine tatsächlichen Frauenkörper zeigten (diese waren vollständig computergeneriert), hat die Technik dennoch das Potenzial, emotionale Schäden zu verursachen. Solche gefälschten Bilder können leicht als echte Aufnahmen wahrgenommen werden und als «Revenge Porn»[4] Verwendung finden. Damit wird nicht nur das Recht am eigenen Bild, sondern auch das Persönlichkeitsrecht verletzt – Deepfakes könnten also auch strafrechtlich relevant werden. Tatsächlich ist diese Art der Persönlichkeitsrechtsverletzung der mit Abstand verbreitetste Anwendungsfall: Dem Sicherheitsunternehmen Deeptrace Labs zufolge machen gefakte Pornos 96 Prozent aller Deepfakes aus[5].

Neben der Möglichkeit des Cybermobbings von Frauen wird befürchtet, dass künftig auch Minderheiten und andere gefährdete Gruppen zu Opfern von Deepfakes werden könnten (wie beispielsweise ethnischen Minderheiten oder LGBTQ-Personen).

Vertrauensverlust

Die wachsende Zahl manipulierter Filme birgt auch eine allgemeinere Gefahr für die Gesellschaft: den Vertrauensverlust, der damit einhergeht. Gezielt gestreute Falschnachrichten haben die Grenzen zwischen Fakten, Desinformation und Lügen aufgeweicht. Deepfakes lassen Menschen zweifeln, ob sie ihren Augen und Ohren trauen können. Selbst bewegte Bilder gelten plötzlich nicht mehr als Beweis, Wahrheit wird relativ.

Wenn jedes Video, jede Tonaufnahme eine Lüge sein kann, wird es für Schuldige einfacher, die Wahrheit als Fälschung abzutun. So mehren sich an den Gerichten die Fälle, in denen behauptet wird, dass Beweisvideos gefälscht seien. Der Effekt, dass man als Zuschauer das Gefühl hat, nichts mehr glauben zu können, kann auch eine Bedrohung für eine Demokratie darstellen. Es genügt dabei, die Glaubwürdigkeit politischer Gegner und deren Aussagen infrage zu stellen und echtes Filmmaterial als Fälschung abzutun.

Ein Beispiel dafür findet sich im afrikanischen Land Gabun: Dessen Präsident Ali Bongo war aufgrund schwerer Krankheit monatelang nicht in der Öffentlichkeit aufgetreten – es traten bereits Gerüchte auf, er sei verstorben. Als dann ein Video auftauchte, in welchem er eine Neujahrsansprache hielt, wurde dieses von seinen politischen Gegnern als Deepfake bezeichnet. Dies

---

[4] Als Racheporno bezeichnet man pornografische bzw. freizügige Videos oder Bilder von einer Person, die ohne deren Einwilligung, oftmals im Rahmen eines Racheaktes veröffentlicht werden.

[5] Ajder, H., Patrini, G., Cavalli, F., Cullen. L. (2019). *The State of Deepfakes: Landscape, Threats, and Impact*. Deeptrace Lab. https://enough.org/objects/Deeptrace-the-State-of-Deepfakes-2019.pdf

löste Unruhen aus, welche letztlich in einem (erfolglosen) Militärputsch endeten. Bis heute gibt es keine Beweise für eine Manipulation – doch allein der Verdacht genügte schon, um das Land in eine politische Krise zu stürzen.

Mögliche Deepfake-Anwendungen und Risiken

Zurzeit werden Deepfakes am häufigsten zu Unterhaltungszwecken eingesetzt. Künftig sind zudem folgende Anwendungen vorstellbar:

- In der Bildung könnten Deepfakes zur besseren Veranschaulichung des Unterrichtinhalts genutzt werden (z.B. Zeitzeugen aufleben lassen).
- In Marketing und PR könnten Deepfakes eingesetzt werden, um Kunden und Kundinnen mit individuell angepassten Botschaften anzusprechen.
- Einzelhändler könnten Kunden und Kundinnen dank Deepfakes anbieten Kleidung virtuell anzuprobieren.
- Menschen, die ihre Stimme durch eine Krankheit oder Unfall verloren haben, könnte die Deepfake-Technologie helfen, ihre eigene Stimme wiederherzustellen (Stimm-Synthese zu medizinischen Zwecken).
- Dank Deepfake könnten künftig neue Bilder im Stil bereits verstorbener Artisten erschaffen werden.
- In der Trauerbewältigung könnten Deepfakes genutzt werden, um sich beispielsweise verabschieden zu können.

In der Debatte um Deepfakes werden jedoch hauptsächlich Risiken und missbräuchliche Anwendungen genannt. Das EPFL International Risk Governance Center sieht dabei folgende drei möglichen negativen Auswirkungen[6]: Reputationsschäden, finanziellen Betrug oder Erpressung sowie Manipulation von Entscheidungsprozessen. Dabei können die negativen Auswirkungen von Deepfakes entweder Individuen betreffen, Organisationen und Institutionen des öffentlichen und privaten Sektors oder die gesamte Gesellschaft (s. Tabelle 1).

[6] Collins, A. (2019). *Forged Authenticity: Governing Deepfake Risks*. Lausanne: EPFL International Risk Governance Center. https://infoscience.epfl.ch/record/273296.

Tabelle 1[7]

Negative Auswirkungen von Deepfakes

	Reputations schaden	**Finanzieller Schaden**	**Manipulation von Entscheidungs-prozessen**
Individuelle Ebene	• Einschüchterung, Beleidigung • Diffamierung	• Identitätsdiebstahl • Phishing-Betrug • Erpressung	• Angriffe auf einzelne Politiker und Politikerinnen
Organisations-ebene	• Markenschaden • Vertrauensverlust in Organisation	• Aktienkurs-manipulation • Versicherungsbetrug	• gefälschte Gerichts-beweise • Medienmanipulation • gefälschte Ausbildungs-unterlagen • Angriffe auf politische Parteien, Lobby-gruppen usw.
Gesellschaft-liche Ebene	• Schädigung des gesellschaftlichen Zusammenhalts, Erosion der gesellschaftlichen Vertrauensbasis usw. • Wahlmanipulation im In- oder Ausland • Bewusstes Schüren von Spannungen/Panik/Konflikten		

Erste Massnahmen

Angesichts der befürchteten negativen Auswirkungen von Deepfakes werden bereits erste Massnahmen dagegen ergriffen. Das chinesische Ministerium für Cyberspace hat bekannt gegeben, dass das Verbreiten von Deepfakes ohne Kennzeichnung in China seit dem 1. Januar 2020 strafbar ist. Mit Blick auf die Präsidentenwahl 2020 verbot auch Kalifornien gefälschte Foto-, Video- und Audioaufnahmen von Politikern und Politikerinnen. Allerdings befürchten Kritiker und Kritikerinnen, dass das Gesetz schwer umzusetzen sein dürfte.

Reddit, Youtube, Facebook und Twitter haben Deepfakes ebenfalls den Kampf angesagt und angekündigt, manipulierte Videos zu kennzeichnen oder sogar zu löschen (mit Ausnahme von Satirebeiträgen und Parodien). Denn insbesondere vor der US-Präsidentschaftswahl im November wuchs die Sorge vor Falschinformationen im Netz, die Wähler manipulieren könnten. Der Druck auf Online-Dienste ist deshalb gross, gegen solche Falschinformationen vorzugehen.

[7] Nach Collins (2019), übersetzt.

Facebook hat zudem eine Deepfake Detection Challenge ins Leben gerufen. Mit dem Wettbewerb soll die Forschungsgemeinde motiviert werden, sich im Kampf gegen Deepfakes zu engagieren. Doch selbst mit einer technologischen Lösung zur Entlarvung von Deepfake-Videos blieben doch einige Probleme bestehen. Das künstlich verlangsamte Video von Nancy Pelosi wurde beispielsweise klar manipuliert, entspricht jedoch nicht einem Deepfake (sondern vielmehr einem sogenannten «Cheapfake») und könnte demnach auch mit einer solchen Software weiterhin kursieren.

Zahlreiche Fragen

Manipulierte Bilder und insbesondere Fake Videos erschüttern mithin kulturell tief verankerte Gewissheiten. Seit einiger Zeit sind sich Nutzerinnen und Nutzer gewohnt, dass nicht alles stimmt, was im Internet steht. Nun scheint dies auch für Videos und Fotos zu gelten. Man scheint also nicht immer seinen Augen trauen zu können.

Aufnahmen von Überwachungskameras wird bisher beispielsweise erhebliche polizeiliche und juristische Relevanz zugesprochen. Doch wieviel Gewicht kann man solchen Videos noch beimessen, wenn sie sich relativ leicht fälschen lassen?

Wie steht es um den Schutz von Fotos, die eine Person von sich selber online veröffentlicht, ohne damit zu rechnen, dass sie später in einem völlig anderen Kontext verfälscht – etwa gar in einem Film – verwendet werden?

Umgekehrt ist aber auch vorstellbar, dass gesellschaftlich geächtete und verbotene Darstellungen in Zukunft gänzlich digital erzeugt werden. Wie steht es um die strafrechtlichen Folgen, wenn eine Person auf ihrem Rechner einen Racheakt am unbeliebten Nachbarn simuliert und das entsprechende Video speichert, das allein aufgrund von Avataren und Videomanipulation entstanden ist? Ist etwas, das in der Realität verwerflich und strafbar ist, auch dann zu ahnden, wenn es als rein «künstliches» Produkt erzeugt wurde und keine reale Person bei der Entstehung missbraucht wurde und leiden musste?

Heute weisen digital «rekonstruierte» Videos oftmals noch kleinere Unstimmigkeiten auf, anhand derer sie sich als «gefakte» Filme erkennen lassen. Der technische Fortschritt dürfte es aber zunehmend erschweren, gefälschtes Videomaterial zu erkennen. Brauchen wir Spezialisten, um den «Videobeweis» zu retten?

2. Interessante Fragestellungen für eine TA-SWISS-Studie

Allgemein

- Wo sind heute überall manipulierte Realitäten anzutreffen? Wie häufig sind wir diesen ausgesetzt? Was sind die potenziellen Chancen und Risiken?
- Welche (technischen) Möglichkeiten existieren, um sowohl echte als auch manipulierte Inhalte zu erkennen und kennzeichnen?
- Sollen Grenzen für die Schaffung von Deepfakes gesetzt werden? Braucht es ethische Vorgaben, Standards und/oder Best Practices für die Entwicklung und Verbreitung von Werkzeugen, die Deepfakes erstellen können? Wer ist dafür zuständig?

Individuelle Ebene

- Welchen Einfluss hat die omnipräsente Darstellung einer «geschönten» Realität auf Personen – insbesondere auf Jugendliche? Was sind die psychologischen Folgen?
- Wie sollen Kinder, Jugendliche und Erwachsene für den Umgang mit manipulierten «Realitäten» geschult werden?
- Wie kann sich jemand gegen die Verwendung seiner/ihrer Bilder und Videos schützen, wie dagegen vorgehen?
- Welche Gesetze bestehen, um Einzelpersonen vor missbräuchlichen Deepfakes (z.B. für Mobbing oder Erpressung) zu schützen? Braucht es weitere Regulierungen, spezifisch für Deepfakes?

Unternehmen, Organisationen

- Wie können sich Unternehmen (wie Banken oder Versicherungen) vor Deepfakes schützen?
- Wie können Medienschaffende sensibilisiert und ausgebildet werden, um echte Inhalte von manipulierten unterscheiden zu können? Sind Veränderungen in der Recherchearbeit nötig?
- Braucht es Anpassungen in der (Schul-)Bildung und Ausbildung von Medienschaffenden?
- Welche Rolle spielen künftig Ton-, Bilder- und Videobeweise vor Gericht?
- Reicht die bestehende Gesetzgebung, um Unternehmen und Organisationen vor missbräuchlichen Deepfakes (z.B. Betrug oder Rufschädigung) zu schützen? Sind weitere Regulierungen nötig?

Gesellschaftliche Ebene

- Welchen Effekt hat die rasante Entwicklung von manipulierten Realitäten auf unser Vertrauen in die Medien und die Gesellschaft? Können wir unseren Augen noch trauen? Müssen wir künftig Medieninhalten misstrauen, bis deren Authentizität zweifellos festgestellt ist? Was sind die möglichen Folgen für die Gesellschaft und die Demokratie?

3. Données sur la teneur et le déroulement de l'étude

3.1. Teneur de l'étude

L'**étude interdisciplinaire** évaluera les effets des **manipulations numériques du son, de l'image et du matériel vidéo ainsi que les deepfakes ou**, en français, **les hypertrucages**. Cela recouvre les images retouchées, dites « embellies », sur les médias sociaux ainsi que les contenus visuels et auditifs créés ou modifiés grâce à l'apprentissage en profondeur (deep learning). L'objectif est de montrer où se cachent les réalités manipulées aujourd'hui et ce à quoi de futures applications pourraient ressembler.

L'étude visera à évaluer les possibilités (techniques) qui existent pour **détecter** les contenus manipulés et les **distinguer** des contenus authentiques. Il s'agira également de clarifier si et comment des limites à la création d'hypertrucages doivent être posées et qui devrait en avoir la responsabilité.

L'étude examinera par ailleurs quelles sont les **conséquences (psychologiques)** des « réalités embellies » et des hypertrucages sur les individus. À cet égard, il sera également intéressant d'analyser comment les particuliers et les entreprises peuvent se protéger contre les abus. Le rôle que les réalités manipulées et les hypertrucages joueront à l'avenir dans la **formation** (écoles, formation d'adultes, formation des professionnels des médias) sera lui aussi évalué.

De même, il s'agira d'étudier les **aspects sociétaux** des réalités manipulées et leur impact sur la confiance dans les médias et la société. Il conviendra de clarifier s'il faut s'attendre à une perte de confiance et quelles en seraient les conséquences potentielles pour la société.

Dans le **contexte juridique**, il sera nécessaire de vérifier dans quelle mesure la législation en vigueur en Suisse protège les personnes et les entreprises contre les hypertrucages abusifs et s'il est nécessaire de compléter la réglementation actuelle.

Enfin, une **évaluation d'ensemble** sera réalisée et servira de base pour tirer des **conclusions** avec, si possible, des **recommandations** sur la manière d'aborder cette problématique à l'intention des décisionnaires, et notamment des responsables politiques.

3.2. Déroulement, calendrier et dépôt des dossiers

Dépôt des esquisses de projet

La mise au concours se déroulera en deux étapes. Dans un premier temps, les esquisses de projet seront déposées. Elles devront comprendre 4 pages maximum et décrire l'approche choisie :

- Introduction (1 page maximum)
- Problématiques, approche prévue et méthodes de recherche (2 pages maximum)
- Composition prévue de l'équipe de recherche (1 page maximum)

Les esquisses de projet doivent être soumises par voie électronique (en format pdf) à info@ta-swiss.ch. La date limite de soumission est fixée au **16 avril 2021**.

La décision concernant le choix des équipes de projet invitées à poursuivre la procédure de soumission tombera, selon toute probabilité, en mai 2021.

Dépôt des propositions détaillées

Sur la base des esquisses de projet, environ trois équipes seront invitées à la deuxième étape de la procédure de soumission. Les équipes de recherche sélectionnées seront informées en juin et seront invitées à soumettre leur proposition détaillée avant le **22 août 2021** dernier délai. Pour la deuxième étape, les propositions doivent satisfaire aux « Règles à suivre pour la présentation des dossiers de candidature » selon le point 4 (pages 14-15) du descriptif détaillé.

3.3. Réalisation de l'étude

Le Secrétariat de la Fondation pour l'évaluation des choix technologiques mettra sur pied un groupe de spécialistes (dit groupe d'accompagnement) représentatif des différents aspects thématiques de l'étude. La proposition acceptée sera présentée à ce groupe d'accompagnement avant que ne débute sa réalisation, lequel pourra, d'entente avec le Secrétariat, influer sur les priorités et la marche à suivre. Pendant la durée de l'étude, le groupe de projet rédigera de trois à cinq documents de travail ou rapports intermédiaires à l'intention du groupe d'accompagnement et du Secrétariat. Ces comptes rendus serviront de base de discussion, étant entendu que chaque nouvelle phase du projet ne sera entreprise qu'avec l'accord de ces deux instances.

3.4. Budget et calendrier

- Cadre budgétaire : CHF 100 000.- à 160 000.-
- Début de la réalisation : octobre 2021 (éventuellement plus tard, à discuter)
- Durée du projet : 12 à 15 mois environ

Dans ce cadre budgétaire, la TVA est incluse ; il incombe au groupe de projet d'examiner son éventuel assujettissement à la TVA.

3.5. Autres dispositions

- TA-SWISS n'est pas soumis au droit des marchés publics. Cela signifie qu'il n'existe pas de voie de recours ordinaire contre des décisions relatives à l'acceptation ou au refus d'esquisses ou de propositions de projets.
- Aucune correspondance ne sera échangée au sujet des esquisses ou propositions de projets déposées.
- Les partenaires contractuels potentiels n'ont droit à aucun dédommagement pour l'élaboration d'esquisses ou de propositions de projets.
- S'appliquent, lors de l'attribution du mandat, les conditions mentionnées dans le *contrat* entre TA-SWISS et les partenaires contractuels ainsi que les *Directives pour les groupes d'accompagnement d'études de TA-SWISS*, jointes au contrat.

4. Règles à suivre pour la présentation des dossiers de candidature

Nous vous prions de structurer votre proposition selon le schéma de soumission suivant (étant entendu que les sous-rubriques ne sont que des **exemples** et peuvent, par conséquent, être adaptées à la spécificité du cas) :

1. Analyse de la situation : positionnement et justification de la recherche
- Raisons justifiant une étude TA sur le thème proposé
- Portée nationale et internationale du sujet
- Enjeux technologiques, économiques, politiques et sociaux
- État des connaissances avec mise en relief des aspects utiles à la TA
- Avancées prévisibles dans le domaine d'investigation envisagé

2. Exposé de la problématique
- Questions auxquelles il s'agit de répondre
- Objectifs concrets de la proposition ou de l'étude
- Nouveaux résultats et nouvelles conceptions amenés par l'étude

3. Structuration et délimitation de la recherche
- Groupes ciblés et points de focalisation
- Eventuellement: subdivision en projet principal et sous-projets
- Liens existants ou prévus avec d'autres projets traitant de problématiques similaires (contacts nationaux et internationaux)

4. Méthodologie
- Méthodes entrant en ligne de compte pour traiter le sujet (élaboration de variantes)
- Évaluation de ces méthodes en fonction de la problématique et arguments en faveur de celle proposée
- Description de la démarche empirique

5. Coordination du projet
- Composition de l'équipe: chef(fe) de projet et collaborateurs(trices)
- Composition du ou des groupes d'experts
- Principales institutions et personnes de contact (partenaires éventuels; voir aussi point 3)

6. Prestations antérieures
- Listage des travaux déjà réalisés dans le domaine concerné par les membres de l'équipe de projet

7. Programme de travail

- Calendrier énumérant les tâches à accomplir avec indication des délais et des dates d'achèvement ainsi que des responsables de leur observation

8. Plan de financement

- Budget prévisionnel détaillé avec évaluation des moyens nécessaires à la réalisation de chacune des tâches (ou phases) telles que définies au point 7.

9. Diffusion des résultats

- Moyens à mettre en œuvre pour informer l'opinion
- Listage des groupes cibles particulièrement visés et des moyens à utiliser pour les atteindre
- Estimation du coût supplémentaire engendré par la diffusion des résultats