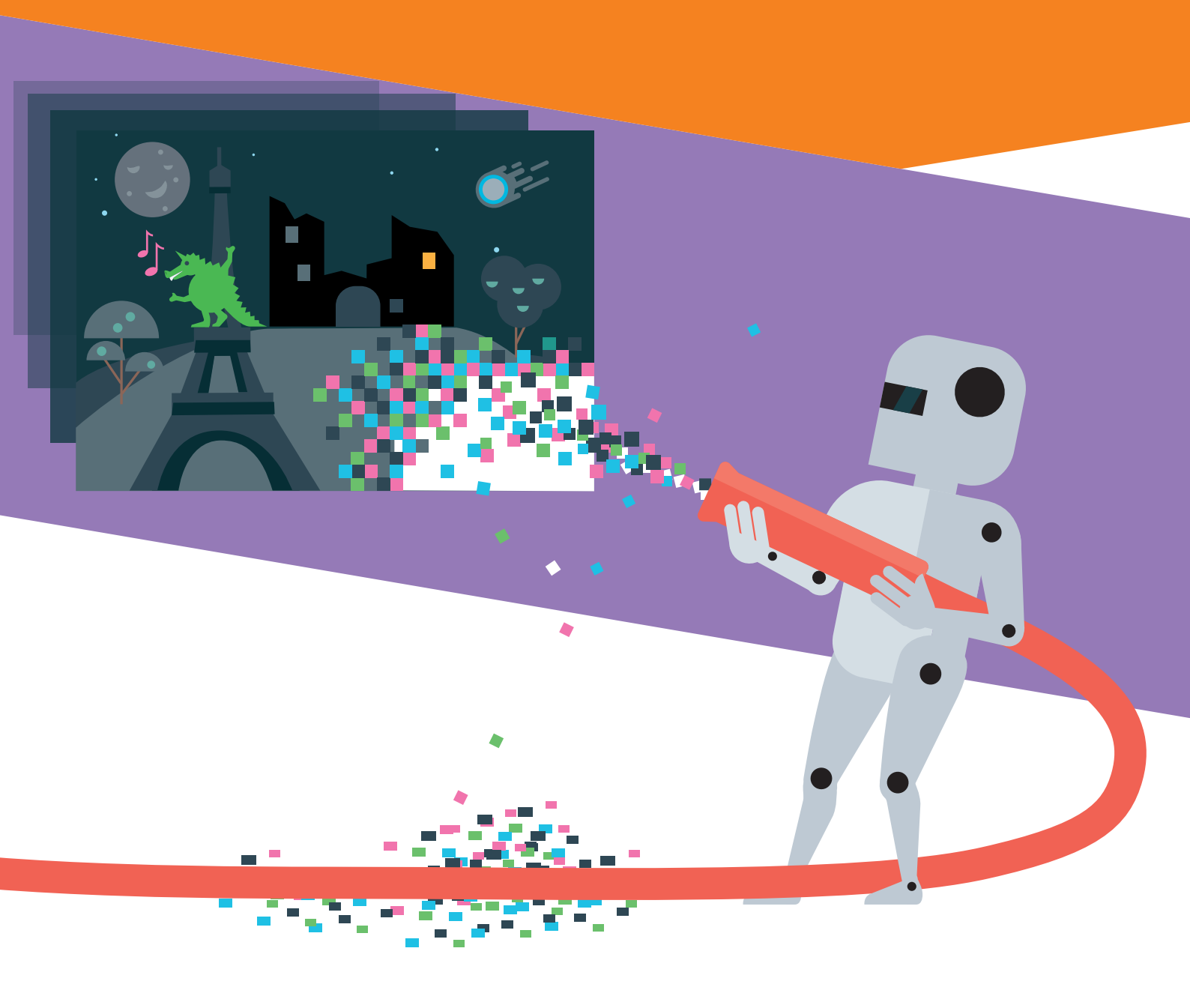
Una sfida per gli occhi e le orecchie

Versione breve dello studio di TA-SWISS «Deepfake e realtà manipolate»



TA-SWISS, Fondazione per la valutazione delle scelte tecnologiche e centro di competenza delle Accademie svizzere delle scienze, intende riflettere sulle ripercussioni – opportunità e rischi – dell'uso di nuove tecnologie.

Questa sintesi si basa su uno studio scientifico effettuato per conto di TA-SWISS da un team di progetto interdisciplinare sotto la guida di Murat Karaboga (Istituto Fraunhofer di ricerca sui sistemi e l'innovazione ISI di Karlsruhe). Vi hanno partecipato Nula Frei (Istituto di diritto europeo presso l'Università di Friborgo), Manuel Puppis e Patric Raemy (Dipartimento di scienze della comunicazione e dei media, Università di Friborgo), Daniel Vogler (Centro di ricerca sulla sfera pubblica e la società fög, Università di Zurigo), Frank Ebbers (Competence Center Nuove tecnologie all'ISI, Karlsruhe), Greta Runge (ISI, Karlsruhe), Adrian Rauchfleisch (Graduate Institute of Journalism presso la National Taiwan University), Gabriele de Seta (Dipartimento di linguistica, letteratura e studi estetici dell'Università di Bergen), Gwendolyn Gurr (Audience Data Analyst, Schweizer Radio und Fernsehen SRF), Michael Friedewald (ISI, Karlsruhe), Sophia Rovelli (Istituto di diritto europeo all'Università di Friborgo).

Questa sintesi presenta i principali risultati e le raccomandazioni dello studio in forma condensata e si rivolge a un pubblico non specializzato.

Versione breve dello studio di TA-SWISS «Deepfake e realtà manipolate»

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

TA-SWISS, Fondazione per la valutazione delle scelte tecnologiche (a cura di).

vdf Hochschulverlag an der ETH Zürich, 2024.

ISBN 978-3-7281-4185-9

Lo studio può essere scaricato gratuitamente: www.vdf.ch

E' disponibile in rete anche questa sintesi: www.ta-swiss.ch

Deepfake: punti essenziali

Lo sviluppo dell'intelligenza artificiale (IA) avanza a grandi passi, seguito a ruota dalla creazione di video, immagini e audio «sintetici», che non riproducono situazioni reali, ma sono generati da programmi informatici. Con tutta probabilità, la crescente facilità con cui si possono produrre questi deepfake ne accrescerà rapidamente il significato sociale. Si delineano opportunità nei settori dell'intrattenimento nonché dell'educazione e della formazione. Non mancano però i rischi – in particolare nei dibattiti politici, in termini di mobbing nei confronti di singoli individui e nell'ambito dei reati economici.

Il termine deepfake – o media sintetici – designa fotografie, video o registrazioni audio prodotte mediante l'intelligenza artificiale, che riproducono una situazione che non è mai esistita nella realtà. Può trattarsi di file manipolati o di file completamente artificiali, generati da software basati su dati di addestramento estratti da enormi raccolte di dati in Internet. I programmi deepfake diffusi al giorno d'oggi comprendono un'ampia gamma di prodotti, da software di facile impiego per la sostituzione dei volti ad applicazioni complesse per creare «messe in scena virtuali» con persone artificiali. Esistono inoltre già primi programmi capaci di generare video – per ora ancora rudimentali – in base a comandi di testo («prompt»).

Alcune opportunità …

I media sintetici offrono potenziali positivi per l'industria dell'intrattenimento. Appaiono promettenti anche altre applicazioni economiche – si possono ad esempio creare influencer artificiali, che presentino indumenti o altri prodotti. Le lezioni di storia nelle scuole potrebbero diventare più appassionanti se avatar di personaggi di epoche passate – Giulio Cesare, Caterina la Grande, Napoleone – potessero interagire con gli allievi. Anche le autorità inquirenti potrebbero approfittare della possibilità di visualizzare il corso degli eventi nell'ambito delle indagini criminali.

… e rischi

I deepfake possono essere usati impropriamente per mostrare persone nel quadro di atti illeciti che non hanno mai commesso o per mettere loro in bocca parole che non hanno mai pronunciato. Simili video o registrazioni audio possono servire a ricattare o a compromettere una persona – un procedimento usato nelle campagne politiche. Possono avere effetti devastanti anche nelle relazioni private, ad esempio sotto forma di «revenge porn» (o pornovendetta).

La voce clonata di una persona può essere usata con intenti fraudolenti per ottenere soldi da amici o familiari. Le voci clonate dei superiori possono invece essere usate impropriamente per altri reati economici, come il furto di segreti commerciali.

I video sintetici rappresentano sfide notevoli per i media, che devono verificarli accuratamente per non contribuire alla diffusione di deepfake.

Raccomandazioni urgenti

Di fronte all'evoluzione rapidissima della tecnica, solo una combinazione di vari provvedimenti di protezione può garantire la possibilità di sfruttare i potenziali positivi dei media sintetici, limitando gli effetti nocivi. Le misure politiche, i rilevatori di deepfake, l'etichettatura dei media sintetici ipotizzata dai grandi fornitori di software e la sensibilizzazione sui deepfake da parte dei media devono completarsi a vicenda. È importante anche la responsabilità individuale del singolo: occorre dar prova di sano scetticismo nei confronti dei video postati su Internet e caricare con moderazione immagini e video privati.

Lo Stato dovrebbe obbligare le piattaforme online a cancellare i deepfake che danneggiano le persone. Siccome nei conflitti con le grandi piattaforme online, in genere i singoli partono svantaggiati, occorrono servizi specializzati che consiglino e sostengano le vittime di deepfake – o coloro che subiscono cancellazioni ingiustificate. La Confederazione e i Cantoni dovrebbero dotare i consultori per le vittime di reati informatici di sufficienti risorse.

Uno sguardo distorto sulla realtà

In genere consideriamo vero ciò che vediamo con i nostri occhi e sentiamo con le nostre orecchie. Desta pochi sospetti in particolare il materiale video, di cui raramente si mette in dubbio la veridicità. Perlomeno fino a qualche anno fa era così. Oggi la tecnica permette invece di produrre facilmente video e registrazioni audio apparentemente veri di fatti che non sono mai avvenuti.

Alla fine dell'autunno del 2023, per un breve periodo sembrava che fosse nata una futura top model: i video postati da Emily Pellegrini nel suo nuovo canale Instagram mandavano in estasi il pubblico. La schiera di follower cresceva a vista d'occhio. E sotto i commenti, alla lunga serie di emoji di cuoricini e fiamme si aggiungevano innumerevoli richieste di contatto. I giornali riferivano che un calciatore tedesco aveva ripetutamente chiesto un appuntamento all'affascinante Emily e che anche un miliardario, una star del tennis nonché altri campioni del mondo sportivo le facevano la corte – fino a doversi arrendere al fatto che non rivaleggiavano per una donna in carne e ossa, ma per un avatar generato mediante l'intelligenza artificiale (IA). A servire da modello all'IA era stata la «donna dei sogni dell'uomo medio»: è quanto aveva reso noto il suo creatore – rimasto anonimo – a cui, stando al britannico Daily Mail, la Emily artificiale aveva fruttato 10 000 dollari al mese.

La «fun-loving girl» (come si autodefiniva) Emily Pellegrini, che mostrava i suoi tratti estetici su piattaforme a pagamento come «Onlyfans» e «Fanvue», è il simbolo di un nuovo tipo di influencer: figure, prevalentemente femminili, create sinteticamente, ma fatte passare per vere, in grado di chattare con il pubblico grazie a generatori di testo basati sull'IA. I vantaggi offerti alla pubblicità dai personaggi artificiali sono incontestabili: una volta creati non richiedono alcun compenso orario, non si stancano mai e si attengono a qualsiasi istruzione.

Inganni profondi

I software basati sull'intelligenza artificiale (o su reti neuronali artificiali) consentono di creare video che mostrano situazioni che in realtà non sono mai esistite. Può trattarsi di filmati di catastrofi naturali o di esplosioni mai avvenute. Oppure di video di personalità note, che dicono o fanno qualcosa che non hanno mai detto o fatto. Un breve filmato creato dal videoartista di Amsterdam Bob de Jong mostra ad esempio l'ex primo ministro olandese Mark Rutte con il doppio mento tremolante mentre intona con sentimento al violino l'assolo di «Stille Nacht». Bob de Jong ha caricato la sua opera sul suo canale YouTube «Diep Nep».

La traduzione in inglese di «Diep Nep» è «deepfake». Questo termine ha preso piede un po' in tutte le lingue per designare video che sembrano autentici, ma che in realtà sono estremamente manipolati o interamente generati al computer. Gli specialisti parlano anche di «media sintetici». La materia prima necessaria: dati ricavati da Internet, in particolare immagini, video e registrazioni audio attinti ai social media o a piattaforme di video. Se Bob de Jong dichiara apertamente che le sue creazioni sono artefatti, per numerosi deepfake – se non addirittura per la maggior parte di essi – l'autore è invece sconosciuto. Nella presente versione breve, le espressioni «deepfake» e «media sintetici» sono utilizzate come sinonimi.

Opere pionieristiche pescate nel torbido

I primi video falsificati sono comparsi nel 2017 su Reddit, una sorta di bacino elettronico di raccolta di contenuti provenienti dai social media. Ad aver caricato i filmini era stato un utente con il nome «DeepFake», che in video pornografici aveva sostituito il volto dell'attrice originale con quello di Emma Watson, Gal Gadot o di altre star del cinema. Poco dopo, un altro utente di Reddit aveva messo a disposizione un software denominato FakeApp, che permetteva a chiunque di creare deepfake. Ciò che prima era riservato ai danarosi studi di Hollywood – la produzione di grafiche computerizzate 3D che richiedevano grandi quantità di calcoli – era così diventato accessibile a tutti coloro che riuscivano a eseguire le cinque facili operazioni prescritte da FakeApp.

Successivamente il darkweb è stato inondato da masse di porno manipolati. Grazie ai progressi tecnici, i video, inizialmente facilmente riconoscibili come falsi a causa della scarsa risoluzione e dei movimenti a scatti, hanno via via assunto un aspetto sempre più realistico. Da tempo, per tali lavori di bricolage non sono più utilizzate solo le immagini di personaggi famosi. Oggi può diventarne vittima chiunque abbia fatto arrabbiare qualcuno. Nel febbraio 2018 in Wikipedia è stata creata la voce «revenge porn», che menziona anche la creazione di immagini nude falsificate.

In base alle stime, anche oggi la maggior parte dei deepfake caricati è costituita da sequenze pornografiche di donne – benché nel frattempo il genere si sia esteso ad altre categorie, in particolare la politica. Le personalità di cui nella rete circolano numerose riprese forniscono infatti particolarmente tanto materiale utilizzabile per creare video falsificati.

In gara per il miglior fake

Una tecnica che ha destato grande attenzione nell'ambito della creazione di deepfake è la cosiddetta rete generativa avversaria (Generative Adversarial Networks, in breve GAN). Si tratta di programmi informatici in grado di generare immagini simili, ma nuove rispetto a un set di dati di addestramento. I programmi sono formati da due elementi, il generatore e il discriminatore, che si contrappongono. Il generatore cerca di generare immagini simili a quelle contenute nel set di addestramento, mentre il discriminatore è concepito per individuare le differenze tra le nuove immagini generate e i dati di addestramento. In questa gara, il generatore e il discriminatore si perfezionano reciprocamente, in modo che alla fine nascono immagini che stilisticamente non si distinguono quasi più dai modelli veri.

L'approccio alternativo consiste nel fare ricorso ad autoencoder, ossia a reti neuronali artificiali in grado di filtrare caratteristiche essenziali da un set di immagini e di trasferirle ad altre immagini.

Trucchi da burattinaio

Le manipolazioni effettuate nel contesto dei deepfake fanno un ampio uso di trucchi – il che non dice tuttavia nulla sull'effetto generato sul pubblico.

Gli interventi possono limitarsi a modificare l'espressione del volto e i movimenti della bocca di una persona, trasferendo la mimica di un attore sulla persona target: gli specialisti chiamano questa nuova messa in scena di un volto «facial reenact-

ment». Questo approccio è utilizzato anche quando una persona sincronizza sé stessa in un video, ossia quando adatta i suoi movimenti della bocca alle sue parole in un'altra lingua. La società HeyGen Labs ha ad esempio sviluppato un programma video basato sull'IA, che registra l'audio di un video, traduce il contenuto e ritrasferisce nel video le parole pronunciate in un'altra lingua – clonando la voce di chi parla e adattando al contempo il movimento delle labbra al parlato. Giornalisti e moderatori, ad esempio, possono così sincronizzare autonomamente le proprie parole.

Un altro tipo di deepfake, il cosiddetto «morphing», consiste nel fondere i tratti del volto di due individui. Questa tecnica è utilizzata soprattutto negli ambienti criminali per falsificare i documenti d'identità in modo che possano essere utilizzati da più persone.

La sostituzione del volto, o «face swapping», è stata impiegata nella fase iniziale dei deepfake, descritta in apertura, per produrre video pornografici falsi. Questa tecnica è utilizzata anche per scopi ludici. Su Internet circolano app gratuite per «rigirare» scene cinematografiche iconiche, sostituendo ad esempio il volto di Leonardo di Caprio o di Kate Winslet con il proprio ritratto.

I software basati sull'IA sono inoltre in grado di costruire immagini completamente nuove di persone che non esistono neanche. Tali ritratti e video creati da generatori di volti trovano impiego come avatar in videogiochi o come interlocutori virtuali nell'assistenza completamente automatizzata alla clientela.

Il software agisce infine come un burattinaio quando modifica le pose e i movimenti di un'intera persona in un video. Questo tipo di deepfake, detto «full body puppetry», è considerato quello più complicato. Attualmente non esiste ancora un pacchetto IA completo, in grado di creare un intero video con animazione vocale e sintetizzazione della voce. Occorre invece combinare tra di loro diversi programmi difficili da usare e in parte a pagamento: la creazione di video deepfake veramente realistici si rivela quindi un'impresa ardua.

Tra non molto, tuttavia, la creazione di video sintetici con un semplice clic dovrebbe risultare a portata di mano: sul modello dei generatori di immagini basati sull'IA, che in base a comandi di testo generano immagini realistiche come se fossero fotografie di un unicorno o una copia falsa di un quadro di Rembrandt, primi programmi basati sull'IA consentono di generare video deepfake partendo da comandi vocali o di testo. È probabile che, in un futuro non lontano, tali software diventeranno accessibili ad ampie cerchie di utenti.

Voci in provetta

È possibile ingannare non solo l'occhio, ma anche l'orecchio – mediante software in grado di clonare la voce e le espressioni tipiche di una persona. Nel 2018 tra gli specialisti è balzato alla ribalta l'esperimento di una società scozzese, che è riuscita a rianimare, perlomeno acusticamente, il presidente degli Stati Uniti assassinato John F. Kennedy: in base al manoscritto e a numerose registrazioni vocali dell'uomo di Stato è stato possibile generare una riproduzione audio del discorso che non aveva potuto pronunciare nell'autunno del 1963 a Dallas, essendo stato vittima di un attentato. Sono stati imitati perfettamente sia l'accento di Boston sia la cadenza di Kennedy.

Negli ultimi anni la tecnica ha fatto ulteriori passi avanti. Per generare un modello del modo di parlare di una persona, sono sufficienti un laptop standard e pochi secondi di un clip audio – ad esempio di una relazione caricata su YouTube. Ha fatto progressi anche la cosiddetta sintesi vocale. Quest'ultima trova impiego nei software che convertono testi scritti in clip audio. Serve a creare automaticamente audiolibri ed è utilizzata anche dai ciechi per far leggere ad alta voce testi scritti.

Strumenti di identificazione dei deepfake

Già solo l'espressione allarmante «deepfake» indica l'obiettivo perseguito da molti video e registrazioni audio sintetici: si tratta di ingannare il pubblico e di influenzarlo a vantaggio del mittente. Tra gli specialisti sono in discussione varie possibilità per gestire le frodi visive e acustiche.

Un approccio consiste nel rendere trasparente la provenienza delle registrazioni. Ciò potrebbe avvenire mediante una firma digitale, apposta sui file video o audio direttamente durante la registrazione. Questa operazione potrebbe essere realizzata abbastanza facilmente utilizzando le blockchain. Tale «impronta digitale» potrebbe tuttavia essere falsificata, seppur con un onere non indifferente. Voci critiche mettono inoltre in guardia dicendo che in questo modo si metterebbe nelle mani di regimi autoritari e servizi segreti uno strumento per scovare whistleblower, attivisti dei diritti umani e giornalisti scomodi. Una firma potrebbe inoltre attestare la provenienza di una registrazione – ma non il fatto che siano effettivamente stati ripresi i fatti completi o solo una parte di essi. Non potrebbero essere identificati neanche i video che mostrano eventi ricreati con attori. Pur essendo tecnicamente «autentici», tali video potrebbero infatti diffondere fatti non corrispondenti al vero. Uno scenario del genere è descritto in una serie televisiva norvegese, in cui un gruppo di partigiani mostra un video del – presunto – assassinio di un primo ministro contestato per consentirgli di «sparire» più facilmente dalla circolazione.

Rilevare le caratteristiche dei falsi

Altri metodi mirano a smascherare i video o le registrazioni audio artificiali in base a determinate caratteristiche. Per quanto riguarda le voci clonate, i pareri sono però perlopiù unanimi: nel frattempo questi deepfake acustici sembrano così reali al punto che è quasi impossibile distinguerli dalla voce di una persona vera – in particolare al telefono. Le autorità di polizia mettono quindi in guardia contro il «trucco del nipote 3.0», che potrebbe funzionare ancora meglio se la vittima pensa di riconoscere la voce di un familiare, che gli chiede soldi.

A loro volta, video falsificati possono tradirsi per determinati artefatti, che magari non sono necessariamente riconosciuti dall'occhio umano, ma vengono smascherati da algoritmi basati sull'IA. Margini mancanti, dissolvenze innaturali, deformazioni e sfocature possono essere segni di un deepfake. Nel frattempo esistono programmi di rilevazione, che pretendono di smascherare i video fake. Nell'ambito dello studio di TA-SWISS sono stati testati due rilevatori gratuiti. I risultati non sono stati soddisfacenti: entrambi i programmi hanno infatti fornito risultati sbagliati. A rivelarsi problematico è stato non da ultimo il fatto che i rilevatori abbiano dichiarato come falsi alcuni video veri – il che potrebbe minare la credibilità dei contenuti originali. È comunque prevedibile che anche gli sviluppatori di software deepfake conoscano le caratteristiche traditrici e facciano del loro meglio per migliorare i loro programmi di conseguenza – è un gioco del gatto e del topo, in cui per ora i falsari hanno una marcia in più.

Mantenere un sano scetticismo

Per ora il buonsenso sembra competere alla pari dei rilevatori tecnici, quando si tratta di riconoscere i deepfake. Una riflessione critica sulla fonte e sui contenuti dei video, vigilanza nei confronti di dettagli come ciocche di capelli, dita o orecchini incongruenti nonché attenzione dinanzi a un comportamento inusuale della persona filmata possono aiutare a smascherare i deepfake. È inoltre possibile allenare la capacità di riconoscere i video falsificati su siti web come Detectfakes.

Un po' più di diffidenza avrebbe perlomeno risparmiato al calciatore tedesco e ad altri ammiratori di farsi abbindolare da Emily Pellegrini. Confrontando diversi video si nota infatti che le proporzioni dell'influencer artificiale variano. Anche le sue risposte sempre di buonumore e a tratti leggermente ossequiose nelle chat avrebbero dovuto destare qualche sospetto. In caso di perfezione surreale e glamour esagerato, un certo scetticismo è sempre d'obbligo.

Esplorazione dei deepfake con un ampio ventaglio di metodi

Oltre a una ricerca approfondita nella letteratura, nell'ambito dello studio di TA-SWISS sui deepfake sono anche state condotte diverse indagini. Mediante un'indagine online, completata da un esperimento online, sono stati sondati le esperienze e il rapporto della popolazione con i deepfake. Il gruppo di progetto ha inoltre condotto interviste e indagini tra operatori dei media, impiegati dell'amministrazione nonché politici per rilevare la loro valutazione dei rischi, delle opportunità, degli effetti dei video e dei contenuti audio falsi. Il gruppo di progetto ha infine testato vari rilevatori di deepfake gratuiti per verificarne la capacità di riconoscere video falsificati.

Come percepiscono i deepfake la popolazione e i professionisti dei media

In Svizzera, finora la popolazione ha avuto pochi contatti con i deepfake. A offrire maggiori probabilità di imbattersi in essi sono piattaforme come YouTube, TikTok e Instagram. Gli intervistati associano i deepfake perlopiù ai rischi e non sono praticamente in grado di distinguere video deepfake fatti bene dai video reali. Anche le maggiori società mediatiche svizzere percepiscono i deepfake soprattutto come un rischio. Sollevando la tematica dei deepfake, i media giornalistici svolgono un ruolo importante nel sensibilizzare la popolazione.

Lo studio di TA-SWISS è la prima indagine completa dedicata alla percezione dei deepfake in Svizzera. Degli oltre 1300 intervistati, poco più della metà ha dichiarato di conoscere il termine «deepfake» – e poco meno della metà di aver già visto un video deepfake. Una piccola minoranza del due-tre per cento ha già avuto esperienze di creazione e diffusione di deepfake. Nel complesso, i risultati dello studio di TA-SWISS mostrano che, in Svizzera, le persone hanno tendenzialmente poca esperienza con le tecnologie deepfake. In questo contesto, classici fattori d'influenza, come l'età, il sesso e la formazione, che di norma hanno un peso nell'ambito dell'acquisizione di nuove tecniche, svolgono un ruolo quasi irrilevante.

La società è più minacciata dell'individuo

La popolazione svizzera percepisce le tecnologie deepfake più come un rischio che non come un'opportunità. Si teme anzitutto che notizie false diffuse come deepfake possano compromettere la fiducia nei media d'informazione svizzeri. È considerato un po' meno virulento il pericolo che i deepfake possano influenzare le votazioni o le elezioni in Svizzera.

Gli intervistati giudicano piuttosto basso il rischio di restare vittima in prima persona di un deepfake. Spicca inoltre il fatto che, rispetto agli uomini, le donne giudicano superiore tale pericolo – un risultato che non sorprende più di tanto visti i numerosi deepfake pornografici.

L'etichetta influenza la percezione delle opportunità

Quanto alle eventuali opportunità che possono essere associate ai deepfake, gli intervistati si mostrano scettici. L'atteggiamento cambia se il termine «deepfake» è sostituito dall'espressione più neutrale «media sintetici». In uno studio preliminare, gli intervistati sono stati suddivisi in due gruppi, a cui sono stati sottoposti questionari differenti. Un gruppo ha ricevuto il questionario con il termine «deepfake». Per il secondo gruppo, in tutte le domande è stata utilizzata la designazione «media sintetici».

È emerso che l'espressione «media sintetici» è meno nota: se circa due terzi degli intervistati avevano già sentito parlare di «deepfake», la designazione «media sintetici» diceva qualcosa solo a poco più di un terzo di loro. La valutazione dei rischi è pressoché analoga con entrambe le «etichette». Per quanto riguarda invece le opportunità, le cose cambiano: rispetto ai deepfake, ai media sintetici sono riconosciute opportunità decisamente migliori quanto al loro impatto nei media e nell'economia. Il modo in cui la società percepisce i benefici di una tecnica dipende quindi non da ultimo dall'etichetta apposta su di essa.

I consigli servono a poco, è utile acquisire familiarità con i nuovi media

Nello studio di TA-SWISS è emerso quanto sia difficile riconoscere i deepfake. In un esperimento, ai partecipanti sono stati mostrati tre video deepfake e tre video reali, chiedendo loro di stimare ogni volta il contenuto di realtà. Per questo esperimento, gli intervistati sono stati nuovamente suddivisi in due gruppi, uno dei quali all'inizio ha ricevuto brevi consigli per riconoscere i deepfake.

La conclusione: gli intervistati sono apparsi molto insicuri nel loro giudizio. Non sono infatti praticamente stati in grado di distinguere i video deepfake fatti bene dai video reali. Inoltre il gruppo che inizialmente aveva ricevuto i consigli per riconoscere i deepfake non ha valutato i video meglio del gruppo che non aveva ricevuto alcun aiuto.

È invece emerso che le esperienze maturate con i social media hanno una correlazione positiva con il riconoscimento dei deepfake. Un'alfabetizzazione mediatica generale contribuisce a non farsi abbindolare dai video falsificati: bisognerebbe quindi avere familiarità non solo con i mass media tradizionali, ma anche sviluppare la capacità di gestire in modo accorto le informazioni provenienti da fonti sconosciute postate sui social media.

Le sfide per il giornalismo

Il riconoscimento dei dati falsi e della disinformazione rientra tra i compiti di base dei professionisti dei media. Video e registrazioni audio realistici al punto da ingannare pongono però sfide supplementari. Siccome il giornalismo deve accompagnare con spirito critico le vicende politiche e contribuire alla formazione delle opinioni e della volontà nell'opinione pubblica, il riconoscimento corretto dei deepfake da parte dei professionisti dei media assume rilievo per l'intera società. Non ridiffondere (involontariamente) deepfake è però anche nell'interesse dei media stessi: la loro credibilità – e di riflesso il modello di affari delle società mediatiche – rischiano infatti di subire un grave danno di reputazione.

I professionisti dei media hanno il compito di verificare l'autenticità dei video (o delle registrazioni audio) nel minor tempo possibile. Queste verifiche sono però complesse. Al contempo molte società mediatiche sono sotto pressione sul piano delle finanze – non tutte possono permettersi personale specializzato per svolgere questa funzione. I contributi giornalistici possono inoltre sensibilizzare il pubblico nei confronti dei deepfake. Se però a questi ultimi è dato molto (troppo) spazio nella copertura giornalistica vi è il pericolo che tra la popolazione si diffonda uno scetticismo esagerato e che aumenti la diffidenza nei confronti dei contenuti mediatici in generale.

I giornalisti sono molti esposti in pubblico. Come evidenziano esperienze fatte in India e negli Stati Uniti, professionisti dei media di spicco possono a loro volta ritrovarsi vittima di deepfake. E benché in genere i giornalisti svizzeri non siano così in vista come alcuni dei loro colleghi stranieri, anche nel nostro Paese molti di loro sono minacciati. I deepfake potrebbero ampliare l'arsenale delle intimidazioni.

Allarme debole nelle redazioni svizzere

L'indagine condotta tra i professionisti dei media svizzeri nell'ambito dello studio di TA-SWISS rivela che il fenomeno dei deepfake è sì percepito nelle redazioni e trattato nella formazione dei giornalisti, ma non è visto come un rischio pressante. I video falsificati sono piuttosto considerati una sottocategoria di disinformazione. Attualmente, i professionisti dei media in Svizzera non temono (non ancora) di ritrovarsi loro stessi vittima di deepfake.

Le redazioni svizzere sono confrontate con video falsificati soprattutto nell'ambito della copertura dall'estero, ad esempio in relazione alla guerra in Ucraina. Qui le redazioni sono chiamate a riconoscere i video falsi per non amplificarne involontariamente la diffusione. In proposito, le redazioni svizzere potrebbero approfittare delle verifiche dell'autenticità dei video da parte dei grandi media stranieri con team di ricerca ben dotati: è questo uno dei risultati emersi dall'indagine. Gli intervistati ritengono che la Svizzera non sia invece nel mirino dei siti di produzione di deepfake, dal momento che i media svizzeri non hanno una grande eco a livello internazionale.

L'indagine tra le società mediatiche svizzere ha però anche mostrato che in particolare i casi complessi pongono requisiti elevati in termini di verifica – e che controlli accurati nelle redazioni non sono sufficienti, ma devono essere completati da campagne di educazione e di sensibilizzazione del pubblico. Non basta infatti che i media verifichino con spirito critico il contenuto di verità delle notizie. Bisogna piuttosto sviluppare una consapevolezza per l'informazione manipolata all'interno dell'intera società.

Fonti attendibili verso diffusori di deepfake

Tra i professionisti dei media intervistati, i deepfake sono considerati perlopiù un rischio. È individuato un certo potenziale tutt'al più nella personalizzazione delle offerte di notizie tramite una moderazione «sintetica» o il ricorso ad avatar nell'ambito delle ricerche.

L'unico beneficio attribuito alle immagini e ai video falsi è il seguente: possono contribuire a rafforzare la posizione dei media giornalistici quale fonte di informazioni affidabile – a condizione che tali media riescano a riconoscere precocemente i deepfake e le altre manipolazioni distinguendosi così da fonti meno affidabili.

Requisiti giuridici differenti per i media giornalistici e le piattaforme online

Per quanto attiene alla diffusione di video falsi assumono rilievo le disposizioni giuridiche applicabili ai media giornalistici e alle piattaforme online. I media tradizionali sono menzionati nella Costituzione federale della Confederazione svizzera: l'articolo 93 prevede ad esempio che la radio e la televisione debbano presentare gli avvenimenti in modo corretto e riflettere adeguatamente la pluralità delle opinioni. Anche la legge federale sulla radiotelevisione stabilisce che, per quanto riguarda le esigenze minime relative al contenuto del programma, i fatti e gli avvenimenti devono essere presentati correttamente «in modo da consentire al pubblico di formarsi una propria opinione. I pareri personali e i commenti devono essere riconoscibili come tali». Vi sono inoltre istanze presso le quali è possibile presentare ricorso se i media violano i principi di correttezza ed equilibrio.

Le piattaforme online, come i social network o i servizi di video sharing che diffondono contenuti creati dagli utenti, non devono invece segnalare i video deepfake. I contenuti diffusi attraverso i social media sono tutelati dalla libertà di opinione e lo Stato può quindi agire solo contro i contenuti manifestamente illeciti. Ai video pornografici può inoltre applicarsi la legge federale sulla protezione dei minori nei settori dei film e dei videogiochi, che obbliga chi diffonde tali contenuti – compresi i servizi di streaming – ad adottare misure di protezione dei minori ed eventualmente a limitare l'accesso a tali video. Far rispettare il diritto svizzero alle offerte provenienti dall'estero è tuttavia una delle maggiori sfide nell'ambito della gestione dei deepfake.

Quando gli avatar fanno politica e agitano l'economia

Dagli scontri bellici alle campagne elettorali: quando il gioco si fa duro, si ricorre ai deepfake per destabilizzare la controparte. Nell'economia, i film fake rientrano nel repertorio della criminalità informatica.

Non sorprende che i deepfake trovino impiego nelle campagne elettorali per screditare gli avversari politici e confondere il pubblico target.

Un deepfake diffuso nel marzo 2022 mostrava il presidente ucraino Zelensky, che in un discorso apparentemente esortava la popolazione del proprio Paese a capitolare di fronte alle forze dell'esercito russo. In Pakistan, all'inizio del 2024 il leader dell'opposizione Imran Khan, detenuto, si rivolgeva ai suoi compatrioti dalla sua cella – e interveniva nella campagna elettorale con un clone di se stesso generato dall'IA. Anche negli Stati Uniti sono stati fatti tentativi per influenzare la campagna elettorale mediante sotterfugi deepfake. A New Hampshire, nel gennaio 2024 i membri del partito democratico hanno così ricevuto una falsa telefonata di Joe Biden. In questa «robo-call», il presidente esortava il proprio interlocutore a non partecipare alle primarie.

Anche in Svizzera i media sintetici hanno già fatto la loro comparsa nell'arena politica. Nell'estate del 2023 ha destato scalpore un manifesto elettorale generato completamente dall'IA, che mostrava in primo piano un'ambulanza bloccata da attivisti per il clima – un evento che non si è mai verificato in questa forma nel nostro Paese. Nell'ottobre dello stesso anno, un video fake mostrava la Consigliera nazionale Sibel Arslan lanciare un appello in contraddizione con i propri valori e i propri obiettivi.

L'umorismo come stratagemma in campagna elettorale

È un fatto che i deepfake possono seminare confusione tra il pubblico e screditare o intimidire personalità attive politicamente o magari carpire loro informazioni riservate. Al di là dall'effetto distruttivo, i video generati sinteticamente hanno tuttavia anche un potenziale politico positivo. Con contributi umoristici potrebbero infatti stimolare il dibattito politico e favorire la formazione delle opinioni. In occasione delle votazioni i deepfake potrebbero essere utilizzati per illustrare in modo comprensibile fatti complessi.

I politici potrebbero inoltre rivolgersi ai loro elettori con deepfake satirici e umoristici, dichiarati in modo trasparente in quanto tali. L'umorismo e l'arguzia si prestano infatti ad ampliare il proprio raggio d'azione e ad attirare su di sé l'attenzione della popolazione.

Auspicata più vigilanza sulla scena politica

Gli autori dello studio di TA-SWISS hanno chiesto a membri del Parlamento svizzero e a collaboratori dell'Amministrazione federale come percepiscono e giudicano i deepfake.

I video fake sono ormai entrati nella quotidianità politica. Una maggioranza degli intervistati ha infatti indicato che i deepfake sono già una tematica di attualità nel proprio lavoro. Le risposte alla domanda se i video falsificati debbano essere considerati più un rischio o eventualmente un'opportunità sono state unanimi: quasi nessuno intravede aspetti positivi, gli intervistati hanno menzionato esclusivamente rischi.

A dominare la scena sono i pericoli per la democrazia svizzera e la fiducia nelle istituzioni del nostro Paese. Tra i rischi rilevanti, gli intervistati hanno menzionato anche quello di ritrovarsi loro stessi vittime o protagonisti di un deepfake, quello di farsi abbindolare o ancora il rischio che un video falso possa compromettere le relazioni internazionali – benché questi eventi siano ritenuti piuttosto improbabili. Gli intervistati erano inoltre concordi nel sostenere che contro i deepfake sono adottate misure di protezione concrete ancora troppo raramente.

Potenziale d'intrattenimento ed educazione

Rispetto alla politica e all'amministrazione, nell'economia l'immagine dei deepfake è meno negativa. L'industria dell'intrattenimento riconosce loro molteplici benefici – ad esempio nell'industria cinematografica. Nell'universo dei videogiochi si cercano possibilità per trasferire il volto dei giocatori sui loro

avatar. E il settore pubblicitario intravede vantaggi negli influencer artificiali, che possono presentare capi d'abbigliamento o svolgere un ruolo attivo nella comunicazione delle imprese.

I deepfake possono anche dare una mano a raccogliere donazioni nelle campagne di istituzioni di utilità pubblica. Nel 2019, una copia sintetica dell'ex star del calcio David Beckham ha lanciato un appello in nove lingue differenti a firmare una petizione per la lotta contro la malaria esortando i leader dei Paesi particolarmente colpiti a impegnarsi maggiormente per contrastare la malattia.

A scuola, l'interesse per le lezioni di storia potrebbe aumentare se Cleopatra, Napoleone o altri personaggi storici interagissero con gli allievi. Avatar potrebbero inoltre accrescere la motivazione dei ragazzi nell'ambito della didattica a distanza personalizzata. Anche la medicina attribuisce ai media sintetici potenziali positivi, ad esempio nell'ambito della terapia contro i disturbi d'ansia inserendo un avatar del curante in una situazione che normalmente genera ansia nel paziente, come il fatto di stare in equilibrio a un'altezza elevata. Dal punto di vista psicologico, ciò permette di analizzare la situazione obiettivamente.

Spionaggio economico con identità rubate

I deepfake non sono tuttavia privi di rischi neanche per l'economia – possono infatti causare danni di reputazione analoghi a quelli che possono verificarsi in politica: esternazioni private falsificate di un presunto insider potrebbero rovinare la reputazione di un'impresa o manipolare i mercati azionari. A loro volta, influencer fake possono essere impiegati per frodi pubblicitarie, compromettendo la fiducia nell'offerente.

I deepfake facilitano inoltre l'usurpazione dell'identità. Una voce clonata o l'avatar tridimensionale di una persona consentono di ingannare i sistemi di riconoscimento della voce e del volto. I criminali possono così accedere al conto di una persona privata o carpire segreti commerciali.

Naturalmente, gli attacchi mediante deepfake nei confronti di attori dell'economia non sollevano interrogativi sostanzialmente nuovi, ma s'iscrivono nel repertorio della criminalità informatica tradizionale. Anche gli obiettivi dei delinquenti informatici e deepfake sono simili: In primo piano vi sono interessi finanziari o il sabotaggio della concorrenza. Sono particolarmente a rischio le imprese o persone che soddisfano quattro criteri: hanno un valore elevato, hanno una visibilità elevata, sono piuttosto lenti nelle loro azioni ed è relativamente facile accedere a essi.

La Svizzera è un bersaglio interessante

Essendo una delle economie più innovative e produttive al mondo, la Svizzera è considerata un bersaglio interessante per i cyberattacchi e quindi anche per gli attacchi mediante deepfake – tanto più che la sua architettura di sicurezza fatica a tenere il passo con il suo ruolo economico di rilievo: secondo il Global Cybersecurity Index 2020, si piazza al 42° posto su 182 Stati. In uno studio condotto lo stesso anno su mandato del Servizio delle attività informative della Confederazione, il 15 per cento delle imprese intervistate ha dichiarato di essere già stata bersaglio di spionaggio economico. Solo una minoranza delle imprese colpite – ossia il 13 per cento – ha però segnalato il caso alla polizia o al ministero pubblico. Molto più spesso sono adottate misure all'interno dell'impresa, spesso con l'ausilio di un sostegno esterno. Il fatto che addirittura ammiraglie dell'economia elvetica non siano immuni ai cyberattacchi è stato dimostrato nel 2023 dagli attacchi contro la Neue Zürcher Zeitung e le FFS.

Siccome molte società tacciono e non denunciano i cyberattacchi, non è possibile quantificare il danno economico causato dai deepfake, tanto più che solo una parte degli attacchi online è attribuito a essi. Esistono invece cifre sul mercato illegale nato attorno ai video falsi: nel 2023 per un semplice video fake nel darkweb bisognava mettere in conto 20 dollari US al minuto. Oltre ai servizi, nelle relative piattaforme si accede anche a istruzioni, contributi alla discussione e altri aiuti che ruotano attorno ai deepfake. Nel darkweb, consigli utili per la distribuzione e l'acquisto di prodotti e servizi legati ai video manipolati sono disponibili soprattutto in forum in lingua inglese e russa, ma registrano un'intensa attività in quest'ambito anche siti darknet in turco, spagnolo e cinese.

Il capitolo conclusivo con le raccomandazioni emerse dallo studio di TA-SWISS illustra le misure che consentono di limitare i rischi rappresentati dai deepfake.

I deepfake agli occhi della legge

In Svizzera, la legge vigente contempla la maggior parte degli illeciti che possono essere commessi mediante deepfake. Far rispettare il diritto è però un'impresa ardua. È necessario puntare sulla collaborazione internazionale.

La libertà dell'arte come pure la libertà di opinione e d'informazione rientrano tra le libertà fondamentali e in Svizzera godono di grande importanza. Sono garantite sia dalla Costituzione federale della Confederazione svizzera sia dalla Convenzione europea per la salvaguardia dei diritti dell'uomo e delle libertà fondamentali. Anche i deepfake beneficiano della tutela dei diritti fondamentali. Questa protezione può tuttavia essere limitata se i contenuti fake violano i diritti degli altri.

Protezione degli autori di prestazioni creative

I media sintetici sono quindi tutelati dal diritto d'autore, sempreché siano considerati un'opera: secondo la legge sul diritto d'autore tra le opere rientrano in particolare le opere fotografiche, cinematografiche e le altre opere visive o audiovisive. Il contenuto artistico o estetico del video o dell'immagine è irrilevante; a essere determinante è il contributo creativo di una persona. In caso di video generato da un'IA in modo interamente automatico manca tuttavia la prestazione creativa. È quindi discutibile se video nati senza alcun intervento dell'uomo debbano essere considerati opere e siano quindi tutelati dal diritto d'autore o dalla libertà dell'arte o meno.

I limiti della libertà d'informazione

Anche la libertà d'informazione ha dei limiti. I dati consapevolmente falsi non sono tutelati. Al di fuori della legge sulla radiotelevisione, che obbliga le emittenti a diffondere un'informazione corretta, non esiste tuttavia alcuna legge che disciplini il contenuto di verità dei video.

Se tuttavia deepfake minacciosi terrorizzassero la popolazione, ad esempio intimidendola annunciando un'imminente catastrofe, gli autori di tali filmati potrebbero essere denunciati penalmente. L'articolo 258 del Codice penale svizzero prevede infatti una pena per chiunque diffonda lo spavento nella popolazione con la minaccia o con il falso annuncio di un pericolo per la vita, la salute o la proprietà.

Furto d'identità, danno di reputazione e frode tramite deepfake

Il Codice civile tutela vari diritti della personalità – tra cui il diritto alla propria immagine, alla propria voce e al proprio nome. I deepfake che utilizzano fotografie e registrazioni audio di una persona senza il suo consenso violano quindi i suoi diritti della personalità.

Chi è messo in scena in modo sconveniente in un deepfake può difendersi: vari articoli del Codice penale puniscono la diffamazione e i delitti contro l'onore. Può appellarsi al Codice penale anche chi viene ingannato con astuzia – ad esempio per indurlo a effettuare un versamento di denaro o a rivelare informazioni riservate.

Il diritto penale punisce la diffusione di contenuti pornografici – una fattispecie adempiuta da buona parte dei deepfake prodotti. È punibile esporre pubblicamente riprese audiovisive pornografiche o offrirle a chi non le ha richieste. Se a una persona viene imposto un deepfake può essere contemplato anche il reato di molestie sessuali. Un nuovo articolo del Codice penale prende di mira il fenomeno della pornovendetta, comminando una pena a chi trasmette a terzi contenuti sessuali non pubblici senza il consenso della persona che vi è riconoscibile.

Sofisticate falsità in documenti

Non di rado, le registrazioni di telecamere di sorveglianza o bodycam trovano impiego come mezzi di prova in procedimenti giudiziari. Tali video potrebbero essere falsificati mediante tecnologie deepfake – o viceversa potrebbero essere utilizzati deepfake per creare falsi alibi. I casi di falsità in documenti sotto forma di documenti manipolati non sono una novità; con i video o le registrazioni audio generati sinteticamente, a essi si aggiunge però un nuovo capitolo.

Di fronte ai deepfake, la valutazione dei mezzi di prova audiovisivi diventa ancora più complessa. In ogni caso, indipendentemente dai mezzi con cui vengono manipolati gli atti processuali, i video o le registrazioni audio, chi falsifica un documento commette un reato. E chi accusa una persona in malafede può a sua volta essere chiamato a rispondere di denuncia mendace o sviamento della giustizia.

I media sintetici quale ausilio nell'ambito del perseguimento penale

Tra gli specialisti del diritto è in corso una discussione sull'uso dei media sintetici quale strumento contro i criminali. Nelle inchieste sotto copertura, spesso bisogna caricare materiale pedopornografico per potersi infiltrare nei corrispondenti forum online. Gli investigatori non possono tuttavia commettere loro stessi reati per dare la caccia ai criminali – e nel nostro Paese la diffusione di pornografia infantile «fittizia» costituisce un reato. Persino la creazione di un deepfake pedopornografico interamente sintetico è problematica, poiché la produzione di un video che sembri realistico richiede dati di addestramento costituiti da immagini di abusi di minori effettivamente commessi. Nelle sue indagini, la polizia non può quindi usare deepfake pedopornografici.

Infine è ipotizzabile impiegare i deepfake nel perseguimento penale – ad esempio per ricostruire un crimine a partire da video di telefoni cellulari, dati di telecamere di sorveglianza e scansioni corporee. Tale procedura solleva però una serie di interrogativi giuridici. A essere problematico è il fatto che, pur sembrando obiettiva, la ricostruzione si basa esclusivamente sulle ipotesi delle autorità inquirenti. Inoltre non è chiaro come si possa concedere all'imputato il diritto di partecipare a tutte le fasi procedurali – compresa l'«ispezione virtuale della scena del crimine» – e come infine debbano essere messe agli atti le prove digitali.

La collaborazione internazionale nella lotta contro i reati globalizzati

La legislazione svizzera contempla quindi la maggior parte dei reati che possono essere commessi con deepfake.

L'applicazione del diritto svizzero rischia tuttavia spesso di fallire per gli ostacoli elevati con cui è confrontata. In genere è infatti difficile identificare gli autori dei deepfake. E anche riuscendo a scovare l'autore c'è poco da fare se il deepfake ha già registrato migliaia di visualizzazioni e raggiunto tutti i continenti. La maggior parte dei deepfake proviene dall'estero ed è caricata su piattaforme al di fuori della Svizzera. Inoltre, in genere i reati commessi in Internet comportano procedimenti complessi. Spesso sono coinvolte più persone, che vanno identificate. A ciò si aggiungono competenze poco chiare e autorità di perseguimento penale sovraccariche di lavoro.

Gli specialisti si aspettano miglioramenti nell'applicazione transfrontaliera del diritto sia dagli accordi di assistenza giudiziaria sia da una cooperazione internazionale approfondita in materia di scambio di dati. L'Unione europea (UE) ha adottato una «normativa sui servizi digitali» (Digital Services Act DSA) volta a migliorare la protezione degli utenti di Internet. La normativa obbliga tra l'altro le piattaforme a lottare contro i contenuti illegali. Le piattaforme devono inoltre consentire agli utenti di segnalare i contenuti e cooperare con cosiddetti «trusted flagger». Questi «segnalatori attendibili» sono istituzioni che scovano e segnalano alle piattaforme i contenuti lesivi. Recentemente, l'UE ha adottato anche un atto giuridico sull'intelligenza artificiale, che prevede un obbligo di trasparenza sui deepfake.

Le piattaforme online stesse, come i social network, non sono rimaste con le mani in mano. Diverse di loro hanno elaborato direttive comuni, che vietano le falsificazioni digitali e le informazioni ingannevoli. Trentaquattro grandi imprese – tra cui figurano Meta, Google, Microsoft e TikTok – hanno inoltre firmato un codice di condotta, con cui s'impegnano a contrastare le informazioni false. Puntare unicamente sull'autodisciplina delle grandi piattaforme non è però sufficiente per tutelare l'interesse pubblico. Alla fine, i criteri di cancellazione sarebbero infatti fissati senza alcuna partecipazione democratica né trasparenza. Il pericolo di un esercizio unilaterale del potere resta reale.

Correggere le distorsioni della realtà: alcune raccomandazioni per gestire i deepfake

Da sole, le misure normative o singole misure tecniche non permettono di evitare e neanche semplicemente di contenere le conseguenze spiacevoli dei deepfake. Occorre piuttosto un mix di provvedimenti a vari livelli e molta responsabilità individuale, per poter approfittare anche del potenziale dei media sintetici.

Siccome la maggior parte dei video manipolati raggiunge il proprio pubblico attraverso le grandi piattaforme online, queste ultime assumono un ruolo chiave nell'ambito del disciplinamento dei deepfake. Sono inoltre chiamati in causa le autorità, il settore della comunicazione, l'educazione – e non da ultimo tutti i cittadini.

Assunzione di responsabilità individuale

La formazione di base e continua sull'alfabetizzazione mediatica e informativa dovrebbe essere in cima all'elenco delle priorità in tutti i settori. A loro volta i cittadini dovrebbero assumersi le loro responsabilità sfruttando le offerte di educazione e informazione dei vari servizi. La responsabilità individuale è irrinunciabile anche nell'ambito della valutazione, della divulgazione e non da ultimo della produzione dei deepfake. Ognuno dovrebbe inoltre essere consapevole del fatto che caricare immagini e registrazioni audio può facilitare la produzione di deepfake. Il principio secondo cui Internet non dimentica vale anche e soprattutto per i deepfake.

Chi guarda volentieri video in Internet o li riceve attraverso i social media dovrebbe sempre considerare la possibilità che si tratti di video falsificati. Occorre dar prova di scetticismo soprattutto nei confronti delle registrazioni che suscitano emozioni o sono particolarmente spettacolari.

Chiamare le piattaforme alle loro responsabilità e rafforzare la protezione delle vittime

Lo Stato dovrebbe adoperarsi per obbligare le piattaforme a cancellare o bloccare i deepfake segnalati. I gestori di piattaforme dovrebbero inoltre essere tenuti a istituire un sistema di segnalazione dei deepfake. Anche requisiti di trasparenza e possibilità di opposizione rafforzerebbero i diritti sia delle vittime dei deepfake sia di coloro che sono colpiti da cancellazioni ingiustificate. Per rispettare tali misure è indispensabile la collaborazione internazionale: occorre quindi creare strumenti supplementari di cooperazione con altri Stati. La Svizzera dovrebbe inoltre impegnarsi affinché siano definite norme e regole applicabili a livello internazionale contro la criminalità informatica.

Nel conflitto con le grandi piattaforme online, in genere il controllo spetta ai singoli. Occorrono quindi servizi specializzati, che consiglino e sostengano le vittime di deepfake – o coloro che sono colpiti da cancellazioni ingiustificate. La Confederazione e i Cantoni dovrebbero dotare i consultori per le vittime specializzati nella criminalità informatica di risorse umane e finanziarie sufficienti. La Svizzera dovrebbe anche riconoscere a livello statale «trusted flagger», che le loro segnalazioni di deepfake problematici in Internet abbiano la priorità; eventualmente andrebbe esaminata l'opportunità di sostenere finanziariamente tali segnalatori.

Sfruttare il progresso tecnico per difendersi

È auspicabile un ampio dibattito sui metodi di autenticazione ed etichettatura. Metodi avanzati, in particolare l'autenticazione a più fattori, possono contribuire a sventare i tentativi di inganno mediante deepfake della voce o del volto. È importante avvalersi, nei limiti delle possibilità, dei metodi di autenticazione più affidabili. I criminali informatici lavorano infatti a loro volta per aggirare le misure di protezione.

Di fronte all'evoluzione rapidissima delle tecniche deepfake è indispensabile utilizzare tutti gli strumenti ipotizzabili per prevenire gli abusi. Anche gli strumenti che al momento sono ancora poco efficaci, come i rilevatori di deepfake, possono costituire un tassello del mosaico di una protezione a tutto campo. È inoltre raccomandabile rendere il più possibile efficaci le misure di sicurezza esistenti.

Educazione sui rischi – e sui benefici

Attualmente solo poche persone hanno già avuto esperienze con i deepfake, mentre molte persone ne sanno poco o niente. Occorre divulgare, attraverso le scuole e i media, informazioni che sensibilizzino sul fenomeno; sarebbero utili in particolare consigli su come verificare le fonti e la plausibilità dei video. Le scuole dovrebbero valutare se l'inserimento dei deepfake negli obiettivi dei piani di studio, come il Lehrplan 21, potrebbe rafforzare l'alfabetizzazione mediatica.

Malgrado i rischi, che vanno spiegati, è importante non soffocare sul nascere i potenziali dei video sintetici. Occorre quindi prestare attenzione a una scelta accurata delle parole. Al termine «deepfake» la gente associa infatti molte meno opportunità che non a espressioni più neutrali come «media sintetici».

Membri del gruppo di accompagnamento

- **Prof. Dr. Reinhard Riedl**, Berner Fachhochschule BFH, presidente del gruppo di accompagnamento, membro del Comitato di direzione di TA-SWISS
- **Dr. Bruno Baeriswyl**, esperto in materia di protezione dei dati, presidente del Comitato di direzione di TA-SWISS
- **Cornelia Diethelm**, Centre for Digital Responsibility
- **Prof. Dr. Rainer Greifeneder**, Abteilung Sozialpsychologie, Universität Basel
- **Thomas Häussler**, Abteilung Medien / Sektion Grundlagen Medien, Ufficio federale delle comunicazioni UFCOM
- **Andrea Hauser**, informatica, Cybersecurity, Scip
- **Erich Herzog**, giurista, Economiesuisse, membro della Direzione
- **Prof. Dr. Selina Ingold**, IDEE Institut für Innovation, Design & Engineering, Ostschweizer Fachhochschule
- **Melanie Kömle Bender**, Mediendokumentalistin, Schweizer Radio und Fernsehen SRF
- **Thomas Müller**, giornalista scientifico, membro del Comitato di direzione di TA-SWISS
- **Prof. Dr. René Schumann**, HES-SO Valais-Wallis, Forschungsinstitut Informatik
- **Prof. Dr. Giatgen Spinas**, Universität Zürich, membro del Comitato di direzione di TA-SWISS
- **Dr. Stefan Vannoni**, economista, CEO cemsuisse, membro del Comitato di direzione di TA-SWISS

Gestione del progetto presso TA-SWISS

- **Dr. Elisabeth Ehrensperger**, Direttrice
- **Dr. Laetitia Ramelet**, Responsabile di progetto
- **Dr. Lucienne Rey**, Responsabile di progetto

Impressum

Una sfida per gli occhi e le orecchie
Sintesi dello studio «Deepfake e realtà manipolate»
TA-SWISS, Berna 2024
TA 81A/2024

Autrice: Lucienne Rey
Traduzione: Giovanna Planzi, Zurigo
Produzione: Laetitia Ramelet e Fabian Schluep, TA-SWISS, Berna
Grafica: Hannes Saxer, Berna
Stampa: Jordi AG – Das Medienhaus, Belp

TA-SWISS – Fondazione per la valutazione delle scelte tecnologiche

Spesso le nuove tecnologie portano netti miglioramenti per la qualità di vita. Talvolta nascondono però anche nuovi rischi, le cui conseguenze non sono sempre prevedibili in anticipo. La fondazione per la valutazione delle scelte tecnologiche TA-SWISS esamina le opportunità e i rischi dei nuovi sviluppi tecnologici in materia di «biotecnologia e medicina», «digitalizzazione e società» e «energia e ambiente». I suoi studi si rivolgono sia ai decisori nella politica e nell'economia che all'opinione pubblica. TA-SWISS promuove inoltre lo scambio di informazioni e opinioni tra specialisti della scienza, dell'economia, della politica e la popolazione attraverso metodi di partecipazione. Siccome devono fornire informazioni il più possibile obiettive, indipendenti e solide sulle opportunità e sui rischi delle nuove tecnologie, i progetti di TA-SWISS sono elaborati d'intesa con gruppi di esperti composti in modo specifico a seconda del tema. Grazie alla competenza dei loro membri, questi gruppi d'accompagnamento coprono un ampio ventaglio di aspetti della tematica esaminata.

La fondazione TA-SWISS è un centro di competenza delle Accademie svizzere delle scienze.

TA-SWISS
Fondazione per la valutazione
delle scelte tecnologiche
Brunngasse 36
CH-3011 Berna
info@ta-swiss.ch
www.ta-swiss.ch

membro delle
accademie svizzere
delle scienze