Deepfake vs. Drepturi digitale

Răspunsuri la întrebări presante legate de un nou orizont informațional.

Din ce în ce mai des vedem pe internet și pe rețelele de socializare videoclipuri cu personalități publice, de obicei din Statele Unite, care par că fac tot felul de lucruri absurde: cântă pe melodii populare, se joacă Minecraft, sau pur și simplu spun lucruri pe care nici un înalt demnitar nu le-ar spune vreodată, în tot felul de contexte ciudate. Pe TikTok, video-uri scurte cu foștii președinți ai Statelor Unite (Obama, Trump și Clinton) jucându-se Minecraft și țipând unul la altul sunt, deja, iconice. Toate acestea sunt făcute posibile cu ajutorul tehnologiilor *deepfake*.

Ce sunt deepfake-urile?

Un deepfake este o imagine sau o înregistrare audio și/sau video artificială (o serie de imagini) generată de un tip special de învățare automată (machine learning/ parte din Inteligența Artificială) numită *deep learning*, de unde provine și numele de *deepfake*. Tehnologia din spatele acestor produse media sintetice (după cum mai sunt numite) este complicată, doar că există o multitudine de metode prin care oricine poate crea un deepfake cu ușurință.

Tehnologiile deepfake recreează și plasează fețele unor persoane - până acum de obicei persoane publice - în contexte inexistente spunând lucruri pe care nu le-au spus, de fapt, niciodată. Un exemplu recent și puternic este video-ul deepfake cu Volodymyr Zelenskyy

care a comandat trupelor ucrainene să se dea bătute în fața rușilor, o situație clară de uz al deepfake-urilor. Deepfake-urile pot fi regăsite peste tot, ipotetic pot fi făcute cu oricine și despre mai orice. Video-uri satirice cu foști președinți Americani adresându-și invective unul celuilalt, au devenit foarte populare pe TikTok, cumulând milioane de vizualizări. Pe TikTok e un adevărat bâlci al deepfake-urilor, ele fiind folosite în cele mai amuzante contexte.

În esență, un deepfake este o înregistrare audio-video sau doar o bucată de imagine sau de audio trucată, făcută să fie ceva ce am putea numi aproape *real* (dar care nu s-a întâmplat niciodată), percepându-l ca fiind plauzibil (ca deepfake-ul viral cu Donald Trump din 2018, cerând Belgiei să se retragă din Acordul de la Paris legat de schimbările climatice), și care imită cu mare precizie gestica, gesturile faciale și vocea persoanei implicate în acel produs.

Cum sunt făcute deepfake-urile?

Nu este foarte ușor să faci un deepfake. Încă. Metoda este destul de tehnică și necesită un computer destul de puternic dacă vrei un produs foarte convingător la final. Primul pas este să treceți mii de fotografii ale fețelor celor două persoane (cea care apare în mod original în video și cea pe care o doriți în locul ei) printr-un algoritm de inteligență artificială numit codificator. Este nevoie de fotografii cu unghiuri cât mai diverse, cu expresii faciale vaste și cu tot felul de mișcări faciale, pentru precizie. Codificatorul găsește și învață asemănările dintre cele două fețe și le reduce la trăsăturile lor comune (expresii faciale, voce, culoarea pielii), comprimând imaginile în acest proces.

Un al doilea algoritm de inteligență artificială, numit decodor, este apoi învățat să recupereze fețele din imaginile comprimate. Deoarece fețele sunt diferite, trebuie antrenat un decodor pentru a recupera fața primei persoane și un alt decodor pentru a recupera fața celei de-a doua persoane. Acest lucru se face manual și necesită mult timp pentru antrenare - lucru pentru care puțini au răbdare, de aceea calitatea multor deepfake-uri de pe internet este foarte joasă. Pentru a efectua schimbul de fețe, este suficient să introduceți imaginile codificate în decodorul "greșit", făcând un schimb între cele două fețe. De exemplu, o imagine comprimată a feței persoanei A este introdusă în decodorul antrenat pe persoana B. Decodorul reconstruiește apoi fața persoanei B cu expresiile și orientarea feței A.

Acesta este doar un mod de a le produce. Mai există și alte metode, un pic mai complicate și mai tehnice decât aceasta, care folosesc inteligența artificială mult mai avansată și complicată, numită Generative Adversarial Networks (un sistem de inteligenta artificiala generativă), care detectează și îmbunătățește orice defecte în deepfake în mai multe runde, făcând mai dificilă decodarea acestora de către detectoarele de deepfake. Probabil în viitor, o dată cu dezvoltările din zona AI, va fi din ce în ce mai simplu.

În orice caz, sărind peste tehnicalități, există metode mult mai ușoare pentru crearea și diseminarea de deepfake-uri. Chiar pentru începători, pot fi folosite aplicațiile următoare: aplicația chineză Zao, DeepFace Lab, FakeApp și Face Swap, sau site-ul deepfakewebs.com,

unde totul e pe sistem de tip *drag and drop*. Momentan, producția de deepfake-uri care să pară cel puțin rezonabile ca aspect necesită mult timp și efort, pe lângă un calculator puternic.

Cum pot fi folosite deepfake-urile pentru încălcarea drepturilor fundamentale?

Între libertate de exprimare și atingeri aduse altor drepturi.

Deepfake-urile sunt niște materiale audio-video trucate, dar care - ca orice conținut informațional - este protejat de libertatea de exprimare. De fapt, modificările unor imagini nu sunt o noutate de multă vreme - doar că au devenit mult mai ușor de făcut și de aplicat la materiale audio-video mai complexe, dar și mai greu de depistat drept imagini trucate.
Pot fi o sumedenie de aplicații extrem de legitime - de la umor, sarcasm, folosire tehnologiei pentru dublarea filmelor, restaurarea vocii persoanelor care au pierdut-o, ori folosirea în zona culturala, care intră în sfera libertății de exprimare. În cele mai multe cazuri de acest tip potențiala problemă a inducerii în eroare se poate rezolva cu un simplu avertisment.

În același timp, deepfake-urile pot fi și sursa unor probleme mult mai grave, în cele mai multe cazuri când sunt rezultatul unei activități intenționate de inducere în eroare. Astfel deepfake-urile sunt o armă potentă pentru crearea de dezinformare și haos informațional pe internet. Multe deepfake-uri sunt folosite pentru creația de pornografie falsă. Așa au și fost botezate produsele media sintetice drept *deepfake*. Totul a început cu un utilizator de Reddit care a lipit fețele unor vedete pe corpurile unor actrițe porno în 2017. Într-un raport, experții în tehnologie de la Deeptrace prezentau situația sumbră: aproximativ 96% din producția de deepfake-uri din 2019 era de pornografie. Și nu orice fel de pornografie, în multe cazuri este vorba de pornografie din răzbunare (revenge porn), unde fețele unor oameni inocenți sunt suprapuse peste filme de natură pornografică cu scopul de a le distruge imaginea publică sau pentru șantaj. Lucrul acesta afectează în mod negativ un număr crescând de oameni - și aproape numai femeile suferă de pe urma acestui tip de uz al deepfake-urilor pentru creația de pornografie. În România, a existat un caz în care un elev de liceu a fost transformat în actor de filme pentru adulți cu ajutorul deepfake-urilor. Povestea următoare este fictivă, dar ilustrativă pentru acest fapt.

Maria are 17 ani și încă este în liceu. Într-o zi, ajungând acasă după o zi lungă de pregătiri pentru bacalaureat, primește un mesaj pe WhatsApp de la un număr pe care nu-l recunoaște. Vede că este un fișier video. Ezită, pentru o secundă, să-l deschidă. Știe că, de obicei, asemenea fișiere nedorite pot să conțină conținut sexual sau pornografic. Zice că mai bine vede ce este, de curiozitate, și deschide fișierul. Toată lumea i se topește în fața ochilor când își vede fața pusă, lipită pe corpul altei fete, într-un film porno. Maria este acum șantajată de un fost partener, amenințată că acest video va fi lansat pe internet dacă ea nu se împacă cu el.

Această poveste scurtă este una fictivă, doar că ilustrează cât se poate de clar natura problemei cu care se pot confrunta și cu care se confruntă unele femei, toate din cauza unor actori rău-intenționați care folosesc o tehnologie puternică. Dar multe femei deja au avut de-a face cu situații similare cu cea prezentată mai sus sau practic tot din aceeași categorie..

Tehnologiile deepfake nu sunt folosite doar pentru pornografie de răzbunare, ci pentru o sumedenie de alte aplicații, de la ilegale la ne-etice, cum ar fi: propagandă politică, studii științifice, șantaj ori fraudă.

Uneori și folosirea pozitivă este suprinzătoare, ca în cazul documentarului de la HBO din 2020 "Welcome to Chechnya" care a folosit tehnologia deepfake pentru a ascunde identitățile refugiaților ruși LGBTQ ale căror vieți erau în pericol, spunându-le în același timp poveștile. Documentarul prezintă un caz fericit al uzului de deepfake-uri și un potențial sănătos în dezvoltarea acestora pentru noi tipuri de artă cu ajutorul tehnologiilor media sintetice

În final, după cum am aratat mai devreme, aceste tehnologii pot fi folosite și în scopuri pozitive, pentru artă, educație, entertainment, sau pentru medicină. Un exemplu, în sensul acesta, este LyreBird, o companie canadiană care folosește tehnologii deep synthesis (un alt tip de deepfake, audio) pentru a clona vocile celor care și-au pierdut vocea, permițându-le să "vorbească" în continuare.

Momentan, adevărul pare unul mai sumbru, și anume că cele mai multe deepfake-uri sunt folosite în scopuri malițioase.

Cum putem depista un deepfake?

Deepfake-urile sunt de obicei sub format audio-video, doar că asta nu este o regulă. Deepfake-urile pot apărea sub orice format disponibil pentru creator (foto, video, audio, și așa mai departe), doar că cele video rămân cele mai convingătoare și puternice arme psihologice de dezinformare, tocmai pentru ca majoritatea utilizatorilor nu se așteaptă că ar putea să fie falsificate. Ele sunt, în același timp, și cele mai populare, fiind mult mai convingătoare decât imaginile false generate cu ajutorul inteligenței artificiale, cum ar fi cele cu fostul președinte al Statelor Unite, Donald Trump, fiind arestat de FBI. Imaginile sunt mult mai ușor de depistat ca fiind false și create cu ajutorul uneltelor care folosesc inteligență artificială, deoarece, în cazul deepfake-urilor, este nevoie de un input uman, care este de un rafinament mult mai înalt decât poate concepe orice unealtă IA generatoare de imagini din zilele noastre.

Video-urile deepfake rămân, pentru moment, cele mai convingătoare produse de aceste gen. Suntem mult mai predispuși la a crede mesajul unui video cu cineva care arată exact ca o ființă umană cunoscută, vorbindu-ne și arătând exact ca orice alt om, decât o fotografie sau doar o bucată de material audio. Pentru o înregistrare video convingătoare, este nevoie de atenție în alegerea fiecărui cadru introdus în program, lucru care necesită multă răbdare și o muncă minuțioasă, atentă la detaliile expresiilor faciale ale persoanelor respective. Pentru că nu mulți creatori de deepfake-uri sunt atât de minuțioși, ele sunt foarte ușor de detectat.

Astfel, deepfake-urile găsite pe rețelele de socializare sunt șubrede și ușor de depistat ca fiind false - ca video-urile cu fostul președinte al Statelor Unite, Michel Obama, cântând piese de artistul Avicii, foarte populare pe TikTok. În asemenea video-uri, fața se mișcă ciudat, gura lui Obama nu se potrivește cu mișcarea pieselor lui Avicii, iar expresiile faciale ale fostului președinte sunt ori inexistente ori prea exagerate.

Astfel de cazuri sunt încă ușor de depistat și demontat, dar creatorii de deepfake și tehnologiile se adaptează și avansează într-un ritm alarmant. Unii cercetători americani au observat că, la fel ca în cazul video-urilor cu Obama, deepfake-urile nu clipesc, făcându-le să arate nerealist și cam neuman. Cum au fost publicate aceste cercetări, cum s-au adaptat creatorii de deepfake și au început să-și facă conținutul fals cu oameni care clipesc. Totuși, există metode prin care putem depista un deepfake cu ușurință, chiar și din cele mai rafinate.

Dave Johnson, scriitor pe zona tech de la Business Insider ne oferă patru întrebări la care ar trebui să găsim un răspuns ca să stabilim dacă ceea ce vedem este un deepfake sau nu:

1) Sunt detaliile clare sau obscure?
2) Cum arată lumina, este clară sau obscură?
3) Se potrivesc și sincronizează cuvintele cu imaginile?
4) Produsul respectiv este dat de o sursă de încredere?

Dacă la toate întrebările (sau la majoritatea dintre ele) astea răspunsul ar fi nu, atunci clar avem de-a face cu un deepfake.

Sunt tehnologiile deepfake periculoase?

Orice tehnologie poate să fie periculoasă dacă este folosită în scop malițios.

În sine, nici o tehnologie nu este periculoasă sau nepericuloasă - totul ține de cum este folosită și cu ce bagaj ideologic, adică set de valori înscrise în tehnologia respectivă, vine la pachet. Cert este că, la momentul actual, tehnologiile deepfake sunt folosite în multe cazuri cu intenții malițioase, pentru șantaj, fraudă sau defăimare. Tehnologiile deepfake trebuie văzute, dacă ar fi să le vedem într-un mod sănătos, ca pe niște unelte, capabile de diverse lucruri și efecte.

Important de ținut în vizor este faptul că, cu cât aceste tehnologii avansează în precizie, cu atât vom vedea mai multe situații în care nu vom putea distinge între ce este real și ce este fals pe internet. Persoane nealfabetizate digital (sau mai puțin alfabetizate), adică persoanele mult mai vulnerabile în mediul online, vor suferi. Un caz deja întâlnit este situația persoanelor mai în vârstă (sau a copiilor), care par a fi sunate de rude ale lor pentru ajutor

financiar de tot felul, dar în spate, de fapt, va fi un actor rău-intenționat cu un deepfake, gată să stoarcă de bani persoane care habar nu au de ceea ce se întâmplă. În trecut (și în prezent), în România, pensionarii erau sunați pe telefoanele lor mobile de tot felul de persoane rău-intenționate care încercau să-i convingă că rudele lor au suferit un accident grav și că e nevoie rapid de o sumă imensă de bani.

Ar trebui sa fie tehnologiile deepfake reglementate?

Cum tehnologia în sine este neutră si depinde de direcția în care este folosită, discuția despre reglementare sau eventuale alte soluții ar trebui făcută mult mai nuanțat. În același timp din experiență de până acum rezultă că anumite folosiri ale tehnologiei ar intra deja sub reglementările existente (fie penal - de ex. infracțiuni de amenințare, șantaj etc. fie civil - utilizare ilegală de date personale, publicitate înșelătoare, etc.)

Aceasta nu înseamnă că nu trebuie să studiem fenomenul și să înțelegem ce și dacă se poate face ceva. Tocmai de aceea unele state (cum ar fi Canada, Uniunea Europeană, Marea Britanie sau Coreea de Sud) discută în primul rând despre cercetare, marcarea conținutului modificat și educație, dar și cum legislația actuală se poate aplica pe anumite tipologii de cazuri. În SUA nu există reglementări federale, dar unele state au legislații care se ocupa de anumite probleme punctuale - de ex. folosirea în campaniile electorale sau diseminarea de deepfake cu conținut sexual explicit. Pe de alta parte China vine cu o abordare tăioasă - în care interzice conținutul deepfake care nu conține o avertizare, și cu obligații dure pentru furnizorii de astfel de materiale.

"Abordarea chinezească" este cea propusă și de un grup de aproape 40 de parlamentari, de la diverse partide (PSD, UDMR, PNL și Minorități) care au propus pe data de 6 aprilie 2023 o propunerea de lege care vizează „interzicerea utilizării malițioase a tehnologiei și limitarea fenomenului Deepfake" propus de aproape 40 de parlamentari, de la diverse partide (PSD, UDMR, PNL și Minorități). Propunerea inițială folosea termenul de *fals sever* pentru deepfake, care era văzut ca un risc la adresa bunăstării societății românești, spunând că astfel de practici sunt făcute în mod deliberat și cu scopul de a înșela. Proiectul inițial conținea și o confuzie este clar făcută, din start, între realitate virtuală (de genul Metaverse) și augmentare a realității (ca în Pokemon GO) de către parlamentari în definiția originală a *deepfake*. Tot varianta inițială avea sancțiuni cu închisoarea de la 6 luni până la 2 ani pentru producția și diseminarea de astfel de creații deepfake

Proiectul a fost adoptat într-o variantă modificată semnificativ de Senat pe 26.06.2023 și acum se afla pe masa Camerei Deputaților. Consiliul Național al Audiovizualului (CNA) este cel care ar trebui să se ocupe de stoparea difuzării de deepfake-uri în media, conform propunerii de lege.

În varianta adoptată de senat, parlamentarii înțeleg prin tehnologii deepfake orice conținut "falsificat de tip imagine, audio și/sau video, realizate, de regulă, cu ajutorul inteligenței artificiale, a realității virtuale (VR), a realității augmentate (AR) sau altor mijloace astfel să creeze aparența că o persoană a spus sau a făcut lucruri, pentru care nu și-a dat consimțământul, care în realitate nu au fost spuse sau făcute de acea persoană".

Producția și diseminarea de astfel de creații deepfake vor putea fi pedepsite cu amenzi de la 10.000 până 100.000 lei. Asta se va întâmpla doar dacă produsul deepfake (audio, video, etc) nu conține un banner care acoperă 10% din suprafața acestuia, pe care să fie specificat faptul că acesta este un produs deepfake.

În cazul produselor deepfake făcute pentru scopuri comerciale, în mod obligatoriu același banner va fi prezent pe care se va menționa faptul ca produsul respectiv conține "ipostaze imaginare". Pentru produsele audio, acestea vor conține o bucată de sunet care să menționeze că este produs deepfake cu ipostaze imaginare.

Rămâne de văzut dacă Camera Deputaților va modifica iarăși semnificativ proiectul.

Multe țări consideră discuția despre reglementarea deepfake-urilor ca fiind ceva de la sine înțeles, fiindcă acestea au un potențial dezastruos de dezinformare a publicului și de destabilizare a democrațiilor consolidate.

Totuși, există un pericol deseori trecut cu vederea - anume reglementarea excesivă a unei tehnologii care are, ca vârsta, doar șase ani (e la clasa 0). Reglementarea excesivă poate fi un impediment pentru inovație în acest mediu al deepfake-urilor, un lucru care, pentru unii, nu este deloc o problemă, fiindcă stopează, din start, potențialul malițios al acestor tehnologii.

O reglementare foarte strictă, una care vede tot ce este deepfake ca fiind, prin natura sa de produs media sintetic, ceva negativ și cu potențial malițios, obturează realitatea posibilității uzului multiplu al acestor tehnologii și potențialul pozitiv pe care astfel de tehnologii îl pot avea, când sunt folosite în scopuri umanitare sau medicinale, artistice și educaționale.

Ultima actualizare - Septembrie 2023