



FRIEDRICH NAUMANN
FOUNDATION For Freedom.

FALSURILE PROFUNDE ȘI DEZINFORMAREA



Agnieszka M. Walorska

ANALIZĂ



**FRIEDRICH NAUMANN
FOUNDATION** For Freedom.

FALSURILE PROFUNDE ȘI DEZINFORMAREA

Agnieszka M. Walorska

ANALIZĂ

IMPRINT

Editura

Fundația Friedrich Naumann pentru Libertate
Karl-Marx-Straße 2
14482 Potsdam
Germania



/freiheit.org



/FriedrichNaumannStiftungFreiheit



/FNFreiheit

Autoare

Agnieszka M. Walorska

Editori

International Department
Global Themes Unit
Fundația Friedrich Naumann pentru Libertate

Concept și aranjare în pagină

TroNa GmbH

Date de contact

Telefon: +49 (0)30 2201 2634

Fax: +49 (0)30 6908 8102

Email: service@freiheit.org

La situația din

mai 2020

Imagini

Photomontages

© Unsplash.de, © freepik.de, P. 30 © AdobeStock

Capturi de ecran

P. 16 © <https://youtu.be/mSalrz8IM1U>

P. 18 © deepnude.to / Agnieszka M. Walorska

P. 19 © thispersondoesnotexist.com

P. 19 © linkedin.com

P. 19 © talktotransformer.com

P. 25 © gltr.io

P. 26 © twitter.com

Restul imaginilor

© Fundația Friedrich Naumann pentru Libertate (Germania)

P. 31 © Agnieszka M. Walorska

Note privind utilizarea acestei publicații

Această publicație este un informativ al Fundației Friedrich Naumann pentru Libertate. Publicația este disponibilă gratuit și nu este destinată comercializării. Aceasta nu poate fi utilizată de către partide sau funcționari electorali în timpul campaniei electorale (alegeri parlamentare, regionale și locale în Germania, și alegeri pentru Parlamentul European).

Licență

Creative Commons (CC BY-NC-ND 4.0)

<https://creativecommons.org/licenses/by-nc-nd/4.0>



CUPRINS

CUPRINS

SUMAR EXECUTIV	6
GLOSAR	9
1.0 STAREA DE DEZVOLTARE INTELIGENȚA ARTIFICIALĂ	
ȘI ROLUL SĂU ÎN RĂSPÂNDIREA DEZINFORMĂRII	15
2.0 FALSURILE IEFTINE ȘI FALSURILE PROFUNDE POSIBILITĂȚILE	
TEHNOLOGICE PENTRU MANIPULAREA TEXTULUI, IMAGINILOR,	
SUNETULUI ȘI IMAGINILOR VIDEO	17
2.1 Falsurile profunde vs falsurile ieftine	18
2.2 Exemple de aplicări	19
Manipularea tiparelor de mișcare	20
Expresiile vocale și faciale	21
Manipularea imaginilor: DeepNude și fețe artificiale	23
Texte generate de IA	24
3.0 DISTRIBUIREA ȘI CONSECINȚELE	24
3.1 Distribuirea	24
3.2 Consecințele	25

3.3 Există exemple de aplicări pozitive ale falsurilor profunde?	27
4.0 COMBATEREA FALSURILOR PROFUNDE	29
4.1 Soluțiile tehnologice pentru identificarea și combaterea falsurilor profunde	29
4.2 Încercările de auto-reglementare de către platformele de socializare	32
4.3 Încercările de reglementare de către legislatori	35
4.4 Responsabilitatea fiecăruia: gândirea critică și competența mediatică	36
5.0 CE URMEAZĂ?	38

SUMAR EXECUTIV



Aplicările Inteligenței Artificiale (IA) joacă un rol tot mai mare în societatea noastră. Noile posibilități ale acestei tehnologii vin însă mână în mână cu noi riscuri. Unul dintre ele este utilizarea abuzivă a tehnologiei pentru diseminarea deliberată a informațiilor false. Deși diseminarea motivată politic a dezinformării cu siguranță nu este un fenomen nou, progresul tehnologic a făcut crearea și distribuirea conținutului manipulat mult mai ușoară și mai eficientă decât înainte. Cu ajutorul algoritmilor IA, imaginile video pot fi acum falsificate rapid și relativ ieftin („deepfakes”), fără să fie nevoie de cunoștințe specializate.

Discuțiile pe această temă s-au concentrat în primul rând pe potențiala utilizare a falsurilor profunde (deepfakes) în campaniile electorale, însă acest tip de imagini video reprezintă doar o mică parte din toate manipulările de acest fel. În 96% din cazuri, falsurile profunde au fost folosite pentru a crea filme pornografice în care apăreau femei celebre. Chiar și femeile din afara sferei publice se pot pomeni în rolul unei vedete involuntare a unor astfel de imagini video manipulate (pornografie deepfake ca act de răzbunare). În plus, aplicațiile precum DeepNude permit transformarea imaginilor statice în imagini nud înșelător de reale. Deloc surprinzător, aceste aplicații funcționează numai dacă în imagini sunt corpuri de femei. Conținutul vizual nu este, însă, singurul tip de conținut care poate fi manipulat sau produs cu ajutorul algoritmilor. Vocile generate de IA au fost deja utilizate cu succes pentru comiterea de fraude care au rezultat în daune financiare mari, iar GPT-2 poate genera texte care inventează fapte și citări arbitrare.

Care este cel mai bun mod de a aborda aceste provocări? Numeroase companii și institute de cercetare au realizat deja un amplu program de investiții în soluții de identificare a imaginilor video generate de IA. Aceste investiții au, de regulă, beneficii de scurtă durată: cei care dezvoltă falsuri profunde răspund la soluțiile de identificare tehnologică cu metode mai sofisticate – un exemplu clasic de cursă a înarmării. Din acest motiv, platformele care distribuie conținut manipulat trebuie să poarte o răspundere mai mare. Facebook și Twitter au în prezent reguli pe care și le-au autoimpus pentru gestionarea conținutului manipulat, însă ele sunt neuniforme și nu se recomandă ca definirea „*libertății de exprimare*” să fie lăsată la discreția companiilor.

Guvernul federal german este, fără îndoială, nepregătit pentru o temă ca „*Aplicațiile conținutului manipulat de IA în scopul dezinformării*”, după cum se arată în succinta investigație parlamentară prezentată de către grupul parlamentar FDP în decembrie 2019. În cadrul guvernului nu au fost clar definite responsabilitățile în raport cu această problemă, iar legi în domeniu nu au fost adoptate. Până acum au fost aplicate doar „*reguli generale și abstracte*”. În răspunsurile

sale, guvernul federal nu sugerează vreo strategie concretă sau vreo intenție de a investi în mai bună pregătire pentru soluționarea acestei probleme. În general, încercările de reglementare existente la nivel german și european par a fi insuficiente pentru a reprimă problema dezinformării la baza căreia stă inteligența artificială. Evident că lucrurile nu trebuie neapărat să stea așa. Unele state din SUA au promulgat deja legi prin care interzic atât pornografia deepfake fără consimțământ, cât și utilizarea acestei tehnologii pentru influențarea alegătorilor.

Respectiv, legislatorii trebuie să elaboreze îndrumări clare pentru platformele digitale în vederea gestionării uniforme a falsurilor profunde, în particular, și a dezinformării, în general. Măsurile pot varia de la etichetarea ca atare și limitarea distribuirii conținutului manipulat (excluzându-l din algoritmii de recomandare), până la ștergerea acestuia. Promovarea competenței mediatice trebuie să fie, de asemenea, o prioritate pentru toți cetățenii, indiferent de vârstă. Este important ca publicul larg să fie informat despre existența falsurilor profunde și ca oamenii să aibă dezvoltată abilitatea de a analiza conținutul audiovizual, chiar dacă devine tot mai greu să se identifice falsurile. În acest sens, merită să luăm în considerare abordarea țărilor nordice, în special Finlanda, a cărei populație a fost considerată cea mai rezistentă la dezinformare.

Există, însă, un lucru pe care nu trebuie să-l facem, și anume să cedăm tentației de a interzice complet falsurile profunde. La fel ca și oricare altă tehnologie, falsurile profunde creează numeroase posibilități interesante, inclusiv pentru educație, film și satiră, în pofida riscurilor pe care le implică.

GLOSAR

Inteligență artificială generală (Artificial General Intelligence) / IA puternică (Strong AI)

Conceptul de IA puternică sau IAG se referă la un sistem informatic care stăpânește o gamă largă de sarcini diferite și, prin urmare, atinge un nivel de inteligență asemănător omului. La momentul de față nu există o astfel de aplicație IA. Spre exemplu, niciun sistem nu poate astăzi să recunoască cancerul, să joace șah și să conducă mașina, chiar dacă există sisteme specializate care pot să îndeplinească fiecare dintre aceste sarcini separat. În prezent, mai multe institute și companii de cercetare lucrează la crearea unei IA puternice, însă nu există vreun consens cu privire la faptul dacă poate fi realizată, iar dacă poate – atunci când.

Big Tech

Termenul „*Big Tech*” este utilizat în mass-media pentru a face referire în mod colectiv la un grup de întreprinderi mari din domeniul tehnologiei informației. Adesea este folosit în mod interschimbabil cu „GAFA” sau „*The Big Four*” pentru a avea în vedere marile companii Google, Apple, Facebook și Amazon (sau „GAFAM” dacă este inclus și Microsoft). Pentru întreprinderile mari chineze din domeniul tehnologiei – Baidu, Alibaba, Tencent și Xiaomi – se utilizează abrevierea BATX.

Falsuri ieftine (Cheapfakes) / superficiale (Shallowfakes)

Spre deosebire de falsurile profunde, falsurile superficiale (shallowfake) sunt manipulări de imagine, sunet sau video create cu tehnologii relativ simple. Exemplele includ reducerea vitezei unei înregistrări audio sau prezentarea conținutului într-un context editat.

GLOSAR

DARPA

Agenția pentru Proiecte de Cercetare Avansate pentru Apărare (Defence Advanced Research Projects Agency) este parte a Departamentului Apărării al SUA, a cărei sarcină este de a cerceta și de a finanța tehnologii militare revoluționare. În trecut, proiectele finanțate de DARPA au avut ca rezultat tehnologii majore care sunt utilizate și în aplicații fără caracter militar, inclusiv internetul, traducerea automată și vehiculele autonome.

Fals profund (Deepfake)

Falsurile profunde (cuvânt telescopat format prin combinarea cuvintelor „deep learning” și „fake”, adică învățare profundă și fals) sunt produsul a doi algoritmi IA care lucrează împreună într-o așa-numită Rețea generativă adversarială (Generative Adversarial Network – GAN). Aceste rețele sunt cel mai bine descrise ca o modalitate de a genera algoritmic noi tipuri de date din seturile de date existente. De exemplu, un GAN ar putea analiza mii de imagini cu Donald Trump, iar apoi ar putea genera o nouă imagine care este similară cu imaginile analizate, dar care nu este o copie identică a nici uneia dintre ele. Această tehnologie poate fi aplicată la numeroase tipuri de conținut: imagini fixe, imagini animate, sunet și text. Termenul „fals profund” este utilizat în primul rând pentru conținut audio și video.

Învățare profundă (Deep Learning)

Învățarea profundă este parte dintr-o familie de metode de învățare automată bazată pe rețele neuronale artificiale care învață din cantități mari de date. La fel ca oamenii care învață din experiență, algoritmi de învățare profundă repetă o sarcină pentru a îmbunătăți treptat rezultatele. Acest proces este numit învățare profundă deoarece rețelele neuronale au mai multe straturi care

fac posibilă învățarea. Învățarea profundă le permite calculatoarelor să rezolve probleme complexe, chiar și atunci când se utilizează seturi de date neuniforme, nestructurate.

Pornografie profundă (Deep Porn)

Pornografia profundă se referă la utilizarea metodelor de învățare profundă pentru generarea de imagini pornografice artificiale.

Rețea generativă adversarială (Generative Adversarial Network)

Rețelele generative adversariale sunt arhitecturi algoritmice bazate pe o pereche de două rețele neuronale, și anume o rețea generativă și alta discriminatorie. Cele două rețele concurează între ele (rețeaua generativă generează date, iar rețeaua discriminatorie falsifică datele) pentru a genera noi seturi de date sintetice. Procesul se repetă de mai multe ori pentru a obține rezultate extrem de similare cu datele reale. Rețelele pot funcționa cu diferite tipuri de date și, prin urmare, pot fi utilizate pentru generarea de imagini, precum și text, sunet sau video.

GPT-2

GPT-2 este un cadru bazat pe o rețea neuronală artificială dezvoltat de compania de cercetare „OpenAI”. Acesta poate genera automat texte în limba engleză. Setul de date pe care se bazează GPT-2 conține în jur de 45 de milioane de pagini de text. Spre deosebire de generatoarele de text tradiționale, GPT-2 nu compune texte din blocuri de text finite și nu se limitează la niciun domeniu specific. Acesta poate genera conținut nou din orice propoziție sau secțiune de text dată.

GLOSAR

IBM Watson

IBM Watson este un sistem bazat pe învățarea automată dezvoltat de către IBM. Prin dezvoltarea lui s-a urmărit crearea unui sistem care poate înțelege și răspunde la întrebări adresate în mod natural. Watson a fost pe larg mediatizat în 2011, când a reușit să-i învingă pe cei mai buni jucători în cadrul concursului televizat „Jeopardy”. De atunci, IBM Watson a fost comercializat ca „IA pentru afaceri”, oferind o gamă de produse cloud și produse de date pentru diferite industrii – de la asistență medicală la producția de filme.

Iarnă AI (AI Winter)

O iarnă AI este o perioadă de finanțare și de interes redus în cercetarea inteligenței artificiale. Termenul a fost inventat prin analogie cu ideea unei ierni nucleare. Ca domeniu tehnologic, IA a cunoscut mai multe cicluri de hype din anii '50, urmate de dezamăgire, critici și reduceri de finanțare.

Rețele neuronale artificiale (Artificial Neural Networks)

Rețelele neuronale artificiale (ANN) sunt sisteme informaționale computerizate inspirate parțial din sistemele nervoase biologice ce se regăsesc în creierul oamenilor și al animalelor. ANN-urile „învață” cum să efectueze sarcini bazate pe exemple, fără a fi programate în baza unor reguli specifice unei sarcini.

ANN-urile pot, de exemplu, să învețe să identifice imagini cu pisici, analizând imagini eșantion care au fost etichetate manual ca „*este pisică*” sau „*nu este pisică*”, iar apoi să folosească rezultatele pentru a identifica pisicile din alte imagini.

Învățare automată (Machine Learning)

În esență, învățarea automată este o metodă care aplică algoritmi pentru a analiza date, pentru a învăța din aceste date, iar apoi pentru a face predicții în

baza lor. Astfel, decât să fie programat manual un software cu instrucțiuni clar definite pentru a efectua o anumită sarcină, software-ul este instruit cu cantități mari de date și algoritmi, care îi permit să învețe cum trebuie efectuată sarcina.

Țintire cu exactitate (Microtargeting)

Țintirea cu exactitate este o tehnică de marketing digital care urmărește să asocieze campaniile publicitare cu consumatorii sau companiile care au cele mai mari șanse să fie interesate de un produs sau serviciu. În funcție de platformă, acest lucru se face, de exemplu, utilizând datele demografice, interesele și istoricul de navigare al publicului țintă. În baza acestor criterii, același organizator de campanie poate aborda în moduri complet diferite destinatari diferiți. Deși a fost dezvoltat inițial pentru campanii politice, acest instrument de marketing este folosit astăzi și în campaniile comerciale.

Phishing

Phishing-ul este o metodă de atac cibernetic cu utilizarea e-mailului. Scopul este de a-l convinge pe destinatarul e-mailului că mesajul este autentic și relevant (de ex. o notificare de la banca sa), pentru a-l motiva să facă clic pe un link sau să descarce un atașament. Anume în acest mod hackerii pot accesa informații sensibile precum parolele.

Pornografie din răzbunare

Pornografia din răzbunare este distribuirea de imagini sau înregistrări video cu conținut sexual ale unor persoane fără consimțământul lor. Făptuitorii sunt de cele mai multe foști parteneri care acționează astfel din răzbunare. Trei sferturi dintre victimele pornografiei ca răzbunare sunt femeile.

GLOSAR

IA slabă (Weak AI) sau IA specializată (Specialised AI)

Algoritmii inteligenței artificiale slabe sunt specializați în îndeplinirea unor sarcini foarte specifice, cum ar fi recunoașterea fețelor, înțelegerea limbajului sau jucarea unei partide de șah. Chiar dacă de regulă sunt mult mai buni sau mai eficienți decât oamenii la aceste sarcini, algoritmii pot doar să rezolve problemele pentru care au fost concepuți. Orice aplicație modernă de IA face parte din categoria inteligenței artificiale slabe, chiar și acele sisteme care par complexe, cum ar fi vehiculele automate sau asistenții lingvistici.

Inginerie socială (Social Engineering)

Ingineria socială se referă la orice măsuri care implică influențarea direcționată a unei persoane sau a mai multor persoane, de exemplu pentru a obține acces la informații confidențiale sau pentru a convinge pe cineva să efectueze o plată. Această practică se mai numește și „*hacking social*” atunci când scopul este de a obține acces la sistemele informatice ale persoanei sau organizației țintă.

Super-inteligență (Super-intelligence)

Super-inteligența este un scenariu ipotetic în care inteligența artificială nu doar că îi întrece pe cei mai inteligenți oameni luați individual, ci și toată inteligența colectivă a omenirii.

Vishing

Vishing-ul (voice phishing) este o metodă de phishing care folosește apelurile telefonice în locul e-mailurilor. Utilizarea falsurilor profunde pentru a genera mesaje vocale poate îmbunătăți eficiența acestei tehnici.



1.0 STAREA DE DEZVOLTARE

INTELIGENȚA ARTIFICIALĂ ȘI ROLUL SĂU ÎN RĂSPÂNDIREA DEZINFORMĂRII

Deși rădăcinile tehnologiei se trag încă de prin mijlocul sec. al XX-lea, inteligența artificială nu a primit prea multă atenție timp îndelungat. Lunga iarnă IA a început să-și reducă din intensitate abia pe la începutul lui 2010. În 2011, sistemul informatic Watson al IBM i-a învins pe cei mai buni jucători în cadrul concursului televizat „Jeopardy”¹, prototipurile Google pentru autovehiculele automate au parcurs peste 100.000 de mile (160.000 km), iar Apple și-a prezentat „asistentul personal inteligent” – Siri. De atunci, interesul public în inteligența artificială, și mai ales în riscurile aferente lor, a crescut constant. Discuțiile despre super-inteligență – declanșate în urma publicării în anul 2014 a unei cărți cu același titlu, scrisă de Nick Bostrom – au generat și mai multă atenție. Personalități de vază au tot continuat de atunci să avertizeze lumea despre pericolele IA, uneori adoptând un ton alarmant.

Stephen Hawking („Dezvoltarea inteligenței artificiale ar însemna sfârșitul rasei umane”) și Elon Musk („AI este un risc existențial fundamental pentru civilizația umană”) sunt frecvent citați. În timp ce super-inteligența și așa-numita „IA puternică” (AGI, Inteligența artificială generală) nu se vor întâmpla prea curând, „IA slabă” și algoritmii săi, probabil, nu atât de slabi, joacă deja un rol în continuă extindere în afaceri, societate și politică. Autorul este de părere că efectele asupra sănătății, energiei, securității, mobilității și multor altor domenii vor fi în mare parte pozitive. Cu toate acestea, ne vom putea bucura de aspectele pozitive ale acestor evoluții numai dacă vor recunoaște și vom contracara cu succes riscurile aferente acestei tehnologii.

Unul dintre ele este utilizarea abuzivă a tehnologiei pentru diseminarea deliberată a informațiilor false. Desigur, dezinformarea motivată politic nu este un fenomen nou. Stalin și Mao sunt cele mai proeminente exemple de dictatori care ordonau cu regularitate să le fie editate fotografiile astfel încât cele vechi

¹ „Jeopardy” este un joc quiz în care participanții primesc indicii generale prezentate ca răspuns, iar ei trebuie să-și formuleze răspunsul sub forma unei întrebări. Între timp, în Germania au apărut concursuri similare precum „Riskant” pe canalul RTL și „Der Große Preis” pe canalul ZDF.



„Cu toate acestea, ne vom putea bucura de aspectele pozitive ale acestor evoluții numai dacă vor recunoaște și vom contracara cu succes riscurile aferente acestei tehnologii.”

să corespundă cu cel mai recent „adevăr”: toți cei care le cădeau din grații erau scoși din imagini, se editau în mod retroactiv noi persoane venite la conducerea partidului, fiind modificat chiar și contextul imaginilor, de exemplu prin schimbarea fundalului. Scopul manipulării acestor înregistrări vizuale a fost de a crea fapte noi, de a rescrie evenimente din trecut și istoria în sine.

Din punct de vedere istoric, efectuarea acestor editări era dificilă și necesita cunoștințe specializate. Astăzi, însă, cu aplicația potrivită pentru smartphone, oricine poate face același lucru fără prea mari eforturi. Tehnologia nu s-a oprit, însă, doar la fotografii. Crearea unor imagini video false, care par adevărate, încă mai necesită eforturi pe măsură. Însă anumite metode de inteligență artificială fac tot mai ușoară manipularea imaginilor video existente. Acestea au devenit cunoscute ca „*falsuri profunde*”. Deși relativ nu sunt atât de frecvente pe internet, utilizarea și diseminarea lor tot mai mare le transformă într-o provocare crescândă pentru societatea noastră. Conținutul manipulat nu doar că se răspândește foarte repede pe platforme precum Facebook sau YouTube, acesta mai vizează în mod special utilizatorii care sunt receptivi la astfel de conținut. Mai mult decât atât, dezinformarea începe să se răspândească tot mai mult și prin intermediul serviciilor de mesagerie precum WhatsApp. În cadrul lor, mesaje criptate sunt distribuite prin conexiuni private, ceea ce crește încrederea în informațiile redirecționate, creând un fel de viralitate ascunsă. Criptarea comunicărilor private online este un deziderat, similar caracterului secret al scrisorilor scrise, aceasta împiedicând vizualizarea mesajelor de către terți. Criptarea, însă, mai înseamnă și imposibilitatea de a verifica acuratețea informațiilor distribuite și de a le modera corespunzător.



2.0 FALSURILE IEFTINE ȘI FALSURILE PROFUNDE

POSSIBILITĂȚILE TEHNOLOGICE PENTRU MANIPULAREA TEXTULUI, IMAGINILOR, SUNETULUI ȘI IMAGINILOR VIDEO

În ultimii doi ani, termenul „fals profund” a devenit tot mai răspândit. Dar ce sunt mai exact falsurile profunde și prin ce se deosebesc de alt conținut manipulat? Deși primele experimente științifice de manipulare a imaginilor video cu ajutorul IA datează pe la sfârșitul anilor '90, publicul larg a aflat despre posibilitățile tehnice abia la sfârșitul anului 2017.

Tot atunci a fost inventată și terminologia, când un utilizator Reddit pe nume „Deepfakes” și alți membri ai comunității Reddit „r/deepfakes” au publicat diferit conținut pe care l-au creat.

Deloc surprinzător, în multe cazuri, falsurile profunde au fost folosite pentru a crea filme porno în care fețele actrițelor erau înlocuite cu fețele unor vedete precum Scarlett Johansson sau Taylor Swift. Un exemplu mai inofensiv a fost filmarea unor scene și înlocuirea feței actorilor cu Nicolas Cage.

„Deloc surprinzător, în multe cazuri, falsurile profunde au fost folosite pentru a crea filme porno în care fețele actrițelor erau înlocuite cu fețele unor vedete precum Scarlett Johansson sau Taylor Swift.”

FALSURILE PROFUNDE FUNCȚIONEAZĂ DUPĂ CUM URMEAZĂ

Falsurile profunde (cuvânt telescopat format prin combinarea cuvintelor „deep learning” și „fake”, adică învățare profundă și fals) sunt produsul a doi algoritmi IA care lucrează împreună într-o așa-numită Rețea generativă adversarială (Generative Adversarial Network – GAN). Rețelele generative adversariale sunt cel mai bine descrise ca o modalitate de a genera algoritmic noi tipuri de date din seturile de date existente.

De exemplu, un GAN ar putea analiza mii de imagini cu Donald Trump, iar apoi ar putea genera o nouă imagine care este similară cu imaginile analizate, dar care nu este o copie identică a nici uneia dintre ele. Această tehnologie poate fi aplicată la numeroase tipuri de conținut: imagini fixe, imagini animate, sunet și text. Termenul „fals profund” este utilizat în primul rând pentru conținut audio și video.

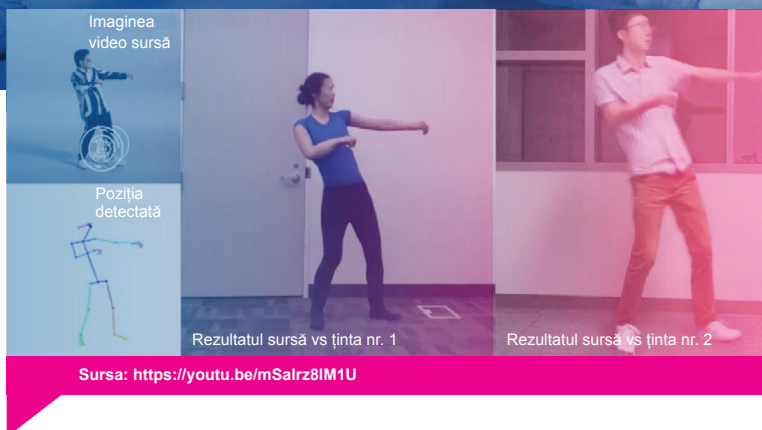
Astăzi sunt necesare doar câteva sute de imagini sau înregistrări audio ca date de antrenament pentru a obține rezultate credibile. Pentru mai puțin de 3 dolari, oricine poate comanda un video fals cu o persoană la alegerea sa, cu condiția să aibă cel puțin 250 de poze cu aceasta – lucru puțin probabil să fie un obstacol pentru orice persoană care are cont pe Instagram sau Facebook. Înregistrările vocale sintetice pot fi, de asemenea, generate pentru doar 10 dolari per 50 de cuvinte.

2.1 Falsurile profunde vs falsurile ieftine

Deși, fără îndoială, sunt cele mai frecvente exemple de falsuri profunde, manipulările pornografice nu sunt în centrul discuțiilor societale actuale. Interesant este faptul că aceste discuții au fost stârnite de un video care nu era nici pe departe un fals profund, ci un simplu fals ieftin (uneori numit și fals superficial): un video cu Președinta Camerei Reprezentanților SUA, Nancy Pelosi, trucat cu mijloace tehnice foarte simple. Înregistrarea a fost încetinită până la aproximativ 75% din viteza inițială. Tonul vocii însă era ridicat, astfel încât să sune natural. Rezultatele: Publicul avea impresia plauzibilă că Nancy Pelosi era beată.

Videoul a fost distribuit de milioane de ori pe rețelele de socializare. Astfel, putem vedea cum chiar și cele mai simple falsuri pot denatura realitatea și pot fi exploatate în scopuri politice. Cu toate acestea, în trecut era foarte greu de falsificat înregistrările astfel încât subiectul să se miște sau să vorbească complet diferit decât în imaginile video originale. Totul până acum.

EXEMPLE DE APLICĂRI



2.2 Manipularea tiparelor de mișcare

În 2018, o aplicație dezvoltată de 4 cercetători din Berkeley a captat extrem de multă atenție. Aceasta utiliza inteligența artificială pentru a transfera dansul obișnuit al unei persoane sursă (de ex. un dansator profesionist) la o persoană țintă.²

Mișcărilor sunt transferate de la videoul sursă la o „*figură schematică*”. Rețeaua neuronală sintetizează apoi videoul țintă în funcție de „*mișcărilor figurii schematice*”.

Rezultatul este un video „*fals*” în care o a treia persoană dansează ca un profesionist. Este clar că acest tip de algoritm ar putea fi folosit nu doar pentru imitarea unor mișcări de dans, ci și pentru generarea unor altor forme de mișcare. Aceasta reprezintă o serie de oportunități pentru portretizarea oponenților politici în situații compromițătoare: Care ar fi, de exemplu, urmările unui video în care un politician ar face un salut nazist sau chiar ar arăta degetul mijlociu?

² <https://arxiv.org/pdf/1808.07371.pdf>

REȚELE NEURONALE ARTIFICIALE

Rețelele neuronale artificiale (ANN) sunt sisteme informaționale computerizate inspirate parțial din sistemele nervoase biologice ce se regăsesc în creierul oamenilor și al animalelor.

ANN-urile „învață” cum să efectueze sarcini bazate pe exemple, fără a fi programate în baza unor reguli specifice unei sarcini. Ele pot, de exemplu, să învețe să identifice imagini cu pisici, analizând imagini eșantion care au fost etichetate manual ca „este pisică” sau „nu este pisică”, și să folosească rezultatele pentru a identifica pisicile din alte imagini.

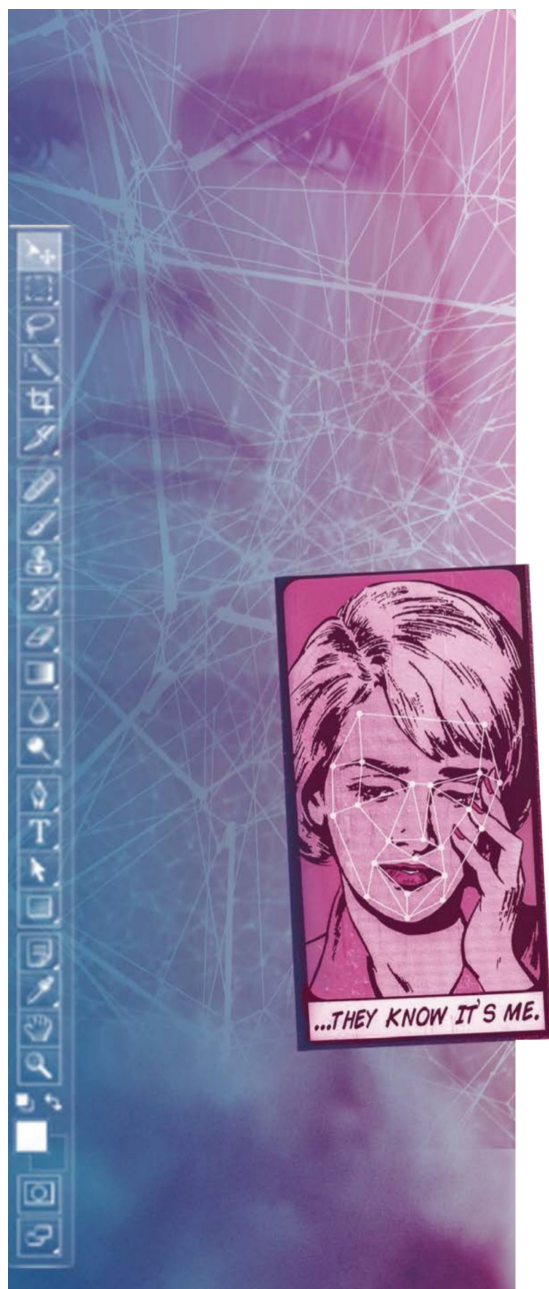
Expresiile vocale și faciale

Falsurile pot avea consecințe și mai ample, făcându-i pe oameni să pară că rostesc cuvinte pe care nu le-au rostit niciodată, acestea fiind însoțite de gesturi, expresii faciale și impresii vocale care par extrem de reale. Mai multe imagini video de acest gen, inclusiv imagini video cu Barack Obama și Mark Zuckerberg, au fost create nu pentru a induce publicul în eroare, ci pentru a demonstra posibilitățile și riscurile acestei tehnologii. Între timp, a existat un caz în care un fals profund a fost creat și distribuit de către o formațiune politică din Belgia – Socialistische Partij Anders (sp.a.).

În luna mai 2018, partidul a publicat pe Facebook un video în care Trump batjocorea autoritățile belgiene pentru respectarea Acordului de la Paris privind schimbările climatice.³

În pofida calității scăzute a imaginilor și a mișcărilor nefirești ale gurii, lucru care ar trebui să trezească suspiciunea oricărei persoane care privește atent, videoul a declanșat sute de comentarii, mulți exprimându-și furia că președintele Americii ar îndrăzni să se amestece în politica Belgiei privind schimbările climatice. Creatorii acestui video, la fel, au încercat să promoveze înțelegerea unei probleme. Videoul era o provocare direcționată spre atragerea atenției

³ <https://www.facebook.com/watch/?v=10155618434657151>



oamenilor asupra unei petiții online în care guvernului belgian i se cerea să întreprindă măsuri mai urgente în legătură cu problemele climatice. Însă ce ar fi fost dacă cineva ar fi creat un video în care Trump ar fi vorbit despre altceva decât politica Belgiei privind schimbările climatice, de exemplu despre intenția sa de a ataca Iranul?

Manipularea imaginilor: DeepNude și fețe artificiale

Conținutul vizual și textual deseori nu este clasificat ca falsuri profunde, chiar dacă poate fi generat cu o tehnologie foarte similară. Există un motiv simplu în acest sens: atât imaginile, cât și textele pot fi manipulate cu atâta ușurință, fără a fi nevoie de o tehnologie complexă, încât „beneficiul” (sau răul, în funcție de perspectivă) acestei manipulari este mult mai mic decât în cazul conținutului audio și video. Pe lângă aceasta, înregistrările video, spre deosebire de conținutul textual și vizual static, sunt mult mai eficiente la declanșarea unor emoții precum frica, furia sau ura.

Cu toate acestea, există exemple de conținut textual/vizual manipulat cu ajutorul IA care au atras, de asemenea, atenția oamenilor. Dacă e să vorbim de imaginile video, scopul primar al

algoritmilor de manipulare a imaginilor este de a crea conținut pornografic fals. Aplicații precum DeepNude pot converti o poză în bikini într-o imagine nud foarte realistă în doar câteva secunde.

Deloc surprinzător, aplicația funcționează doar cu femeile (orice încercare de a selecta o poză cu un bărbat pur și simplu generează organe genitale feminine). Această aplicație, însă, o expune pe fiecare femeie riscului de a deveni o potențială victimă a „pornografiei din răzbunare”, chiar și atunci când nu au existat niciodată poze nud reale cu ea.

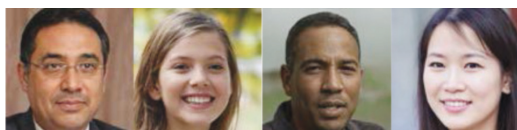


O imagine nud generată cu aplicația deepnude.to

Aceste rețele neuronale nu se opresc doar la manipularea imaginilor cu persoane reale. Ele pot „crea” și persoane total noi sau, cel puțin, fețe total noi.

Aplicațiile comerciale ale acestei tehnologii sunt evidente: bazele de date cu imagini pot fi populate mult mai rentabil cu ajutorul IA decât cu ajutorul unor persoane

reale. Însă acest lucru mai înseamnă și că oricine poate acum să creeze mult mai ușor profiluri false pe rețelele de socializare, de exemplu pentru a distribui conținut politic.



Fețe generate aleatoriu de thispersondoesnotexist.com

Au fost suspectate și încercări de spionaj cu poze de profil generate de calculator, de exemplu profilul pe LinkedIn al unei persoane pe nume „Katie Jones”, o presupusă cercetătoare care lucra la un centru analitic din SUA.

Înainte ca expertiza să identifice mai multe anomalități vizuale ce sugerau faptul că imaginea era sintetizată, profilul a reușit să se conecteze cu 52 de figuri politice din Washington, inclusiv cu un subsecretar adjunct de stat, un consilier principal al unui senator și un proeminent economist.⁴

⁴ <https://www.cnet.com/news/spy-reportedly-used-ai-generated-photo-to-connect-with-targets-on-linkedin/>

Profilul a fost eliminat rapid de LinkedIn, dar se crede că ar fi fost al unei rețele de profiluri fantomă, dintre care unele ar putea să mai existe și care ar putea fi folosite pentru atacuri phishing.

Texte generate de IA

Aplicația descrisă mai sus poate fi implementată eficient mai ales în combinație cu generarea de text cu ajutorul IA.

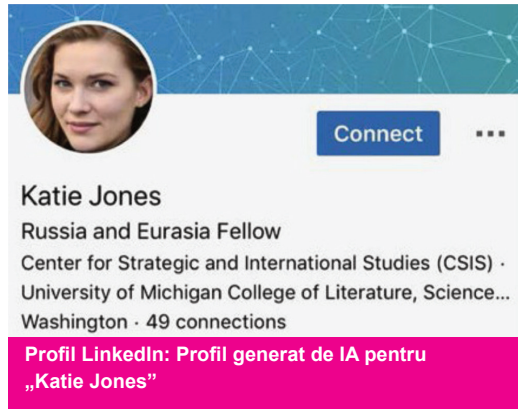
Mulți poate că au auzit deja de această posibilitate, pe care o prezintă generatorul de text GPT-2 creat de compania de cercetare „OpenAI”. Din cauza că ar crea premisele unor abuzuri, GPT-2 a fost la început considerat prea periculos ca să fie pus la dispoziția publicului larg.⁵ Întrucât creatorii săi nu au reușit până acum să găsească nicio dovadă clară că aplicația ar putea fi utilizată în mod abuziv, compania a decis ulterior să publice GPT-2 pe etape.⁶

Chiar dacă încă nu au fost cazuri de utilizare abuzivă, creatorii recunosc faptul că oamenii ar putea, în mare parte, să perceapă textul generat de GPT-2 ca fiind credibil, că generatorul ar putea fi ajustat pentru a produce conținut extremist și că identificarea textului generat ar fi dificilă. Cu aplicația „Talk To Transformer”, oricine poate încerca GPT-2 pentru sine.

Introducerea uneia sau mai multor propoziții în generator are ca output o secvență de text care începe cu inputul propus. Rezultatele sunt deseori, dar nu întotdeauna, surprinzător de coerente. Ele adoptă același ton ca și inputul și simulează credibilitatea inventând experți, statistici și citate.

⁵ <https://openai.com/blog/better-language-models/>

⁶ <https://openai.com/blog/gpt-2-1-5b-release/>



AI-generated fake content could unleash a virtual arms race of misinformation online, experts say.

"Once you get the person to click on something, you've gotten them to put themselves in a position to think a certain way, if they haven't already done so," said Katherine Jellison, a professor at Georgia Tech's School of Interactive Computing and author of the book "Cyberbullying in the Age of the Internet."

Textul cu bold – input-ul, restul textului – generat de IA cu aplicația www.talktotransformer.de

3.0 DISTRIBUIREA ȘI CONSECINȚELE

CÂT DE PERICULOASE SUNT FALSURILE PROFUNDE ÎN REALITATE?

3.1 Distribuirea

Cuantificarea cu exactitate a distribuirii de falsuri profunde este dificilă, mai ales din momentul în care numărul lor este, fără îndoială, în continuă creștere.

Deeptrace, o companie care oferă o soluție tehnologică pentru detectarea falsurilor profunde, a încercat să prezinte o estimare precisă în raportul pe care l-a elaborat: „The State of Deepfakes: Landscape, Threats, and Impact” [Situția falsurilor profunde: prezentare, amenințări și impact].⁷

Potrivit estimărilor pe care le-au făcut autorii raportului, publicat în septembrie 2019, numărul falsurilor profunde aproape că s-a dublat în decursul a 7 luni, de la 7.964 în decembrie 2018 la 14.678 în iulie 2019. Dintre acestea, 96% era conținut pornografic fără consimțământ, în care apăreau exclusiv corpuri de femei.

Principalele victime erau femeile celebre, pe internet putând fi găsite mii de poze false cu ele. Potrivit raportului Deeptrace, numai cele mai populare 4 site-uri web de pornografie profundă luate aparte au înregistrat peste 134 de milioane de vizualizări ale videoclipurilor false în care apăreau femeile celebre. De fenomenul pornografiei din răz bunare menționat mai sus pot fi afectați, însă, și foarte mulți oameni simpli. Creșterea fenomenului este determinată în primul rând de accesibilitatea mai mare la instrumente și servicii care permit crearea de falsuri profunde fără cunoș tințe în domeniul programării.

Și în anul 2019 au fost raportate cazuri de utilizare a clonelor lingvistice generate de IA pentru ingineria socială. În luna august, „The Wall Street Journal”

⁷ The State of Deepfakes: Landscape, Threats, and Impact, Henry Ajder, Giorgio Patrini, Francesco Cavalli și Laurence Cullen, septembrie 2019.

a raportat⁸ despre primul caz de fraudă vocală bazată pe IA – cunoscută și sub denumirea de vishing (prescurtare pentru „voce phishing”) – în valoare de 220.000 de euro pentru compania germană vizată.

Software-ul a imitat vocea managerului din Germania atât de bine, inclusiv intonațiile sale și accentul său ușor german, încât colegul său din Marea Britanie a îndeplinit imediat cererea urgentă a apelantului de a transfera suma indicată. Deși la momentul actual acest incident este unul singular, se pare că în viitor vor exista mai multe încercări de acest fel.

În foarte multe mediatizări ale falsurilor profunde s-a pus accentul pe potențialul lor de a discredita opoziției politici și de a submina procesele democratice. Acest potențial nu s-a materializat până în prezent. Deși au existat imagini video manipulate tehnic ale unor politicieni precum Barack Obama, Donald Trump și Matteo Renzi, acestea au fost motivate în primul rând de satiră sau au fost create în scopuri demonstrative, iar caracterul fals al acestora a fost dezvăluit foarte repede.

3.2 Consecințele

Cu toate acestea, faptul că politicienii încă nu au folosit falsurile profunde în scopuri de dezinformare nu înseamnă că ultimele nu au fost utilizate până acum pentru a influența discursul politic. Un exemplu care a nu a fost mediatizat prea mult în presa occidentală demonstrează modul în care simpla cunoaștere a existenței falsurilor profunde poate afecta climatul politic.

Președintele Gabonului, Ali Bongo, nu a apărut în public mai multe luni după ce a suferit un accident vascular cerebral. Deloc surprinzător, au început să se răspândească zvonuri precum că acesta ar fi decedat. Pentru a pune capăt speculațiilor, președintele a publicat un video în decembrie 2018 cu discursul său obișnuit din noaptea dintre ani. Înregistrarea, însă, a avut efectul opus. Mulți au crezut că Bongo arăta ciudat și imediat au început să suspecteze că imaginile video erau false. La scurt timp, armata a lansat o lovitură de stat fără succes, citând presupusul fals profund drept unul dintre motivele lor.⁹

Expertiza ulterioară, totuși, a confirmat caracterul autentic al înregistrării. Ali Bongo și-a revenit între timp în urma accidentului vascular cerebral și rămâne în funcție.

⁸ <https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceo-voice-in-unusual-cybercrime-case-11567157402>

⁹ <https://www.technologyreview.com/s/614526/the-biggest-threat-of-Deepfakes-isnt-the-Deep-fakes-themselves/>



„Deși au existat imagini video manipulate tehnic ale unor politicieni precum Barack Obama, Donald Trump și Matteo Renzi, acestea au fost motivate în primul rând de satiră sau au fost create în scopuri demonstrative, iar caracterul fals al acestora a fost dezvăluit foarte repede.”

Acest exemplu ne demonstrează faptul că cea mai mare amenințare pe care o prezintă falsurile profunde nu sunt falsurile profunde în sine. Simplul fapt că astfel de imagini video sunt posibile din punct de vedere tehnic ridică următoarea întrebare: Mai putem oare avea încredere în autenticitatea imaginilor video?

Această întrebare va pune la îndoială alegerile prezidențiale din SUA ce au loc în anul 2020. În campania electorală din 2016, dezinformarea și manipularea prin intermediul IA, cel mai proeminent sub formă de țintire cu exactitate și bot-uri, începuseră deja să joace un rol. Falsurile profunde reprezintă acum un alt instrument în arsenalul dezinformării. Chiar dacă în campania electorală sunt folosite puține falsuri profunde sau chiar deloc nu sunt folosite, există probabilitatea ca mulți politicieni să accepte cu recunoștință posibilitatea de a respinge înregistrările reale, dar nefavorabile, ca fiind falsuri.



3.3 Există exemple de aplicări pozitive ale falsurilor profunde?

„Tehnologia ne oferă permanent modalități de a face rău și bine. Ambele iau amploare. [...] Însă faptul că și noi avem o nouă alegere de fiecare dată este un lucru pozitiv”,¹⁰ punctează Kevin Kelly, redactor-șef de ani buni și membru fondator al revistei de tehnologie „Wired”. Ar fi această declarație valabilă și pentru falsurile profunde?

Tehnologia este mai cu seamă promițătoare pentru industria cinematografică, în special în post-producție și dublare. De ce? În prezent, modificarea retroactivă a unei secvențe de dialog este foarte costisitoare pentru studiourile de film. Actorii, echipa de filmare și platoul de filmare trebuie din nou rezervate. Tehnologia din spatele falsurilor profunde ar face posibile aceste editări foarte repede și la costuri foarte mici.

Și dublarea filmelor se poate bucura de îmbunătățiri semnificative. Ar deveni posibil ca mișcarea buzelor actorilor să fie adaptată la cuvintele dublate, iar vocile lor ar putea fi sintetizate pentru a le adapta la limba țintă, ceea ce înseamnă că nu mai este nevoie de dublare.

Un exemplu de astfel de aplicare este o înregistrare video cu David

¹⁰ Citat luat de la https://www.edge.org/conversation/kevin_kelly-the-technium/



„Tehnologia este mai cu seamă promițătoare pentru industria cinematografică, în special în post-producție și dublare.”

Beckham, care promovează o campanie împotriva malariei.¹¹ El „vorbește” în mai multe limbi, iar gura sa pare să se sincronizeze perfect cu cuvintele în fiecare limbă.

Educația este un alt domeniu interesant de aplicare. Un exemplu în acest sens ar putea fi crearea de înregistrări video cu personaje istorice care să-și relateze viața sau să răspundă la întrebări. Foarte multă atenție mediatică a atras proiectul „*Dimensions of History*” (*Dimensiunile istoriei*)¹² al Fundației Shoah a Universității din California de Sud. Este vorba de interviuri și înregistrări holografice a 15 supraviețuitori ai holocaustului. Această expoziție itinerantă a fost expusă în mai multe muzee din SUA, iar recent a fost găzduită de Muzeul suedez de istorie.

Persoanele care au mers la expoziție au avut ocazia să le adreseze întrebări hologramelor. Software-ul pentru recunoașterea vocii le-a adaptat apoi întrebarea cu o secvență din interviu. Cu ajutorul tehnologiei din spatele falsurilor profunde, această utilizarea ar putea fi extinsă la o scală mai largă și în mai multe limbi.

¹¹ <https://www.malariamustdie.com/>

¹² <https://sfi.usc.edu/dit>

4.0 COMBATEREA FALSURILOR PROFUNDE

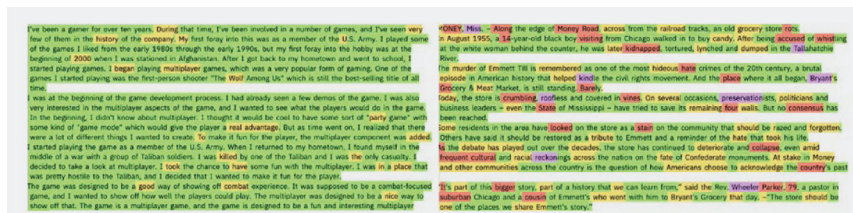
CUM PUTEM FACE FAȚĂ PROVOCĂRILOR ASOCIATE CU FALSURILE PROFUNDE?

Aceste exemple pozitive nu sunt, desigur, menite să reducă la minimum potențialele pericole pe care le prezintă falsurile profunde. Riscurile sunt indiscutabile și necesită contramăsuri decisive – lucru în privința căruia există un consens. Un consens nu atât de puternic există, însă, în privința naturii exacte a acestor contramăsuri. De asemenea, apare întrebarea cum să garantăm drepturile fiecărei persoane la libertatea de exprimare fără a submina nevoia societății pentru un sistem informațional de încredere.

4.1 Soluțiile tehnologice pentru identificarea și combaterea falsurilor profunde

O soluție pentru combaterea falsurilor este dezvoltarea unor tehnologii care să poată deosebi conținutul fals de cel real. Această abordare folosește algoritmi similari cu cei care au generat falsurile pentru prima dată. Cu ajutorul GLTR, un model care are la bază sistemul GPT-2 menționat mai sus, cercetătorii de la Laboratorul de IA al MIT-IBM Watson și HarvardNLP au investigat dacă aceeași tehnologie utilizată pentru a scrie articole fabricate în mod independent poate fi utilizată pentru a recunoaște pasajele de text generate de IA. La generarea unui pasaj de text în aplicația de testare, cuvintele sunt evidențiate cu verde, galben, roșu sau violet pentru a le indica nivelul de previzibilitate, în ordine descrescătoare.

Cu cât este mai mare proporția cuvintelor cu previzibilitate redusă, mai ales secțiunile de text marcate cu roșu și violet, cu atât este mai mare probabilitatea ca autorul pasajului să fie o ființă umană. Cu cât mai mare este previzibilitatea cuvintelor (și textul „mai verde”), cu atât este mai mare probabilitatea ca textul să fie generat automat.



Rezultatele analizei: autor uman vs generator de text, Sursa: gltr.io

Tehnici similare ar putea fi folosite pentru identificarea imaginilor video manipulate. În anul 2018, cercetătorii au observat că actorii din înregistrările video false nu clipeau. Acest lucru se întâmplă din cauză că în imaginile statice utilizate pentru generarea imaginilor video erau prezente de la început persoane cu ochii deschiși. Utilitatea acestei observații, însă, nu a durat mult. De îndată ce aceste informații au devenit publice, au început să apară imagini video cu oameni care clipeau. La aceeași tendință se poate aștepta și în cazul oricăror alte mecanisme de identificare care vor fi descoperite ulterior. Acest joc de-a șoarecele și pisica se întâmplă deja de zeci de ani în domeniul securității cibernetice – progresul este întotdeauna în beneficiul ambelor părți.

Aceasta, însă, nu înseamnă că eforturile de identificare a falsurilor profunde trebuie încetate. În septembrie 2019, Facebook, împreună cu inițiativa PIA¹³, Microsoft și mai multe universități, a anunțat o campanie „Deepfake Detection Challenge”¹⁴ cu un premiu de 10 milioane de dolari.

De asemenea, Facebook a comandat un set de date cu imagini și înregistrări video de la actori care s-au filmat special pentru acest scop, astfel încât provocarea să fie asigurată cu un set de date corespunzătoare cu care să se poată lucra. Câteva săptămâni mai târziu, Google a lansat, de asemenea, un set de date cu 3.000 de înregistrări video manipulate în același scop.

Agentia americană de finanțare în domeniul cercetării DARPA, de asemenea, desfășoară din anul 2016 activități de recunoașterea conținutului manipulat ca parte a programului MediFor (prescurtare pentru Media Forensics), investind peste 68 de milioane de dolari pe parcursul a 2 ani.¹⁵ La momentul actual nu se

¹³ Parteneriatul pentru IA (PIA) este o organizație care reunește universități, cercetători, ONG-uri și întreprinderi pentru a înțelege mai bine impacturile și efectelor IA asupra societății. www.partnershiponai.org

¹⁴ <https://ai.facebook.com/blog/deepfake-detection-challenge/>

¹⁵ <https://futurism.com/darpa-68-million-technology-Deepfakes>

cunosc prea multe dacă în Germania și Europa sunt dezvoltate soluții tehnice pentru combaterea falsurilor profunde, iar dacă sunt dezvoltate – despre ce fel de soluții este vorba.

Majoritatea măsurilor sunt întreprinse de companii aparte, precum compania Deeprtrace menționată mai sus, precum și de proiecte de cercetare precum „Face2Face” de Matthias Nießner¹⁶, profesor la Universitatea Tehnică din München.

Potrivit răspunsului pe care guvernul german l-a oferit la o întrebare adresată în plenul Parlamentului de către grupul parlamentar FDP, „Centrul Național de Cercetare pentru Securitate Cibernetică Aplicată” CRISP/ATHENE lucrează în prezent la această problemă împreună cu Universitatea Tehnică din München și Institutul Fraunhofer.

În plus, radiodifuzorul internațional german Deutsche Welle (DW), Institutul Fraunhofer pentru tehnologia media digitală (IDMT) și Centrul tehnologic din Atena (ATC) au inițiat proiectul comun de cercetare „Digger”. Scopul acestuia este extinderea platformei de verificare web „Truly Media” de către DW și ATC cu tehnologie audio criminalistică de către IDMT Fraunhofer, inter alia, pentru a le oferi asistență jurnaliștilor.¹⁷ Cu toate acestea, răspunsul guvernului federal nu sugerează vreo strategie concretă sau vreo intenție de a investi în soluționarea acestei probleme.

„Cu cât este mai mare proporția cuvintelor cu previzibilitate redusă, mai ales secțiunile de text marcate cu roșu și violet, cu atât este mai mare probabilitatea ca autorul pasajului să fie o ființă umană. Cu cât mai mare este previzibilitatea cuvintelor (și textul „mai verde”), cu atât este mai mare probabilitatea ca textul să fie generat automat.”

¹⁶ <https://niessnerlab.org/projects/thies2016face.html>

¹⁷ <https://dip21.bundestag.de/dip21/btd/19/156/1915657.pdf>



4.2 Încercările de auto-reglementare de către platformele de socializare

Chiar dacă marile întreprinderi din domeniul tehnologiei au contribuit cu date și resurse financiare pentru conceperea unei soluții tehnologice la această problemă, apelurile îndreptate spre Facebook și alte companii similare în vederea luării unor măsuri suplimentare s-au intensificat în condițiile în care platformele acestor companii joacă un rol central în răspândirea dezinformării. Ca răspuns, Twitter și Facebook au publicat, la sfârșitul anului 2019 și, respectiv, la începutul anului 2020, declarații referitoare la planurile lor de a combate fenomenul falsurilor profunde.

În noiembrie 2019, Twitter le-a cerut utilizatorilor săi un feedback cu referire la o „*propunere de politică pentru conținutul media sintetizat și manipulat*”. Orientările au fost apoi anunțate la începutul lunii februarie 2020: orice fotografie, înregistrare audio sau video, care a fost „*editată sau falsificată în mod semnificativ*” cu scopul de a-i induce în eroare pe oameni, va fi eliminată în cazul în care Twitter va considera că aceasta ar putea provoca daune grave, de exemplu, punând în pericol securitatea fizică a persoanelor sau cauzând „*tulburări civile pe scară largă*”. În caz contrar, tweet-urile pot fi etichetate în continuare drept conținut media manipulat: partajarea lor va fi

Is the media significantly and deceptively altered or fabricated?	Is the media shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
☒	☐	☐	Content may be labeled
☒	☐	☒	Content is likely to be labeled, or may be removed.
☒	☒	☐	Content is likely to be labeled.
☒	☒	☒	Content is very likely to be removed.

Twitter: o nouă abordare față de conținutul media sintetizat și manipulat:
https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

însoțită de un avertisment, iar în fluxurile utilizatorilor acest conținut nu va mai avea prioritate. Aceste modificări urmau să intre în vigoare pe 5 mai 2020.¹⁸

Facebook face un pas înainte. Pe 6 ianuarie 2020, Monika Bickert, vicepreședintă pentru managementul politicilor globale în cadrul Facebook, a scris într-o postare pe un blog că falsurile profunde care îndeplinesc anumite criterii vor fi șterse de acum înainte de pe platformă.¹⁹ Potrivit postării, toate conținuturile editate sau sintetizate cu ajutorul IA în moduri care nu sunt evidente pentru o persoană obișnuită vor fi șterse. Pe de altă parte, conținutul satiric este exclus din aceste orientări, ceea ce lasă loc semnificativ pentru interpretare.

Interesant este faptul că aceste orientări nu se aplică și falsurilor ieftine; acestea vizează în mod explicit și exclusiv conținutul generat de IA. Prin urmare, imaginile video false cu Nancy Pelosi, pe care le-am menționat anterior, pot în continuare să fie găsite pe Facebook.²⁰ Chiar dacă a recunoscut că verificatorii săi de fapte au semnalat imaginile respective ca fiind false, Facebook a refuzat să le șteargă, deoarece compania „nu are în vigoare o politică prin care să impună veridicitatea informațiilor postate pe Facebook”.²¹

¹⁸ https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

¹⁹ <https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/>

²⁰ YouTube, pe de altă parte, o altă platformă care contribuie la răspândirea rapidă a informațiilor false prin algoritmi săi de recomandare, a șters videoul, însă a refuzat să facă o declarație clară despre modul în care va gestiona ulterior falsurile profunde.

²¹ <https://www.politico.com/story/2019/05/24/facebook-fake-pelosi-video-1472413>



„La momentul actual nu se cunosc prea multe dacă în Germania și Europa sunt dezvoltate soluții tehnice pentru combaterea falsurilor profunde, iar dacă sunt dezvoltate – despre ce fel de soluții este vorba.”

Această abordare reflectă poziția Facebook cu privire la libertatea de exprimare și depășește problema falsurilor profunde. În cadrul discuțiilor privind publicitatea politică, Rob Leathern, Directorul pentru managementul de produse al Facebook, a scris într-o postare pe blog în ianuarie 2020 că aceste tipuri de decizii nu ar trebui luate de companiile private, *„motiv pentru care pledăm pentru o reglementare care să fie aplicată întregii industrii. În lipsa reglementării, Facebook și alte companii sunt libere să-și aleagă propriile politici”*.

Cu siguranță merită să discutăm dacă modul în care Facebook interpretează libertatea de exprimare este justificată din punct de vedere etic. Cu toate acestea, declarația lui Rob Leathern atrage atenția asupra unei întrebări specifice, și anume lipsa sau cel puțin caracterul incomplet al reglementării.

4.3 Încercările de reglementare de către legislatori

În Germania, falsurile profunde sunt incluse în categoria „*regulilor generale și abstracte*”, potrivit răspunsului guvernului federal la succinta investigație parlamentară prezentată de către grupul parlamentar FDP, după cum am menționat mai sus. „*Nu există reglementări specifice la nivel federal care să acopere exclusiv domeniul de aplicare a falsurilor profunde sau care să fi fost create pentru astfel de aplicări. Guvernul federal revizuieste în permanență cadrul legal la nivel federal pentru a vedea dacă sunt necesare ajustări pentru a face față provocărilor tehnologice sau sociale.*”

Aceasta presupune că unele aspecte parțiale ale problemei falsurilor profunde, printre care pornografia din răzbunare, sunt acoperite în mod implicit de legislația în vigoare, însă de fapt nu există o abordare explicită față de manipularea conținutului modificat. Acest lucru se aplică întregului spectru de dezinformare în spațiul digital, nu doar „*falsurilor profunde*” în mod special. După cum a remarcat autorul studiului „*Regulatory responses to disinformation*” [Răspunsuri de reglementare la dezinformare]²² din Stiftung Neue Verantwortung: „*Încercările anterioare de reglementare și soluțiile politice [în Germania și Europa] nu sunt deloc potrivite pentru reducerea gradului de dezinformare*”. În studiul său „*Deepfake Legislation: A Nationwide Survey*” [Legislația în domeniul falsurilor profunde. Studiu la nivel național],²³ firma de avocatură WilmerHale oferă o analiză detaliată a statutului reglementării falsurilor profunde în SUA.

În legislația penală a Statelor Unite au fost deja introduse legi explicite privind falsurile profunde. În Virginia, de exemplu, se pedepsește pornografia deepfake fără consimțământ, iar în Texas – orice fals profund al cărui scop este influențarea alegătorilor. O lege similară a fost adoptată și în California în septembrie 2019.

Posibil cea mai amănunțită reglementare a falsurilor profunde a fost întreprinsă de legislatorii chinezi la sfârșitul anului 2019. În China, prestatorii și utilizatorii de mesaje video și servicii de informații audio online sunt impuși prin lege să marcheze clar tot conținutul care a fost creat sau editat cu ajutorul noilor tehnologii, cum ar fi inteligența artificială.

Deși merită cu siguranță să analizăm dacă și alte țări ar putea adopta reglementări similare, cazul Chinei ne lasă un gust neplăcut: însuși guvernul chinez

²² https://www.stiftung-nv.de/sites/default/files/regulatorische_reaktionen_auf_desinformation.pdf

²³ Matthew Ferraro, WilmerHale | Deepfake Legislation: A Nationwide Survey – State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media.

folosește dezinformarea bazată pe tehnologie pentru a ținti protestatarii din Hong Kong, inter alia, iar folosirea acestor noi reglementări ca pretext pentru o cenzură ulterioară pare inevitabilă.

Reglementarea eficientă a noilor fenomene tehnologice cu siguranță nu este ușoară. În trecut acest lucru deseori s-a dovedit a fi dificil. De exemplu, pentru a conduce o mașină în Anglia sec. al XIX-lea, era nevoie ca o a doua persoană să meargă în fața acestuia fluturând un steag roșu conform prevederilor Legii din 1865 privind locomotivele.²⁴ Cu toate acestea, există măsuri pe care legislatorii le pot lua deja pentru a contracara fenomenul falsurilor profunde. Întrucât o proporție de 96% dintre falsurile profunde sunt astăzi cazuri de pornografie fără consimțământ, ar fi o idee bună să se înceapă pedepsirea acestora în mod explicit, așa cum s-a făcut în statele Virginia și California. Defăimarea, fraudă și drepturile privind confidențialitatea pot fi reglementate în mod similar. Mai mult decât atât, legislatorii trebuie să elaboreze îndrumări clare pentru platformele digitale în vederea gestionării uniforme a falsurilor profunde, în particular, și a dezinformării, în general.

Aceste măsuri pot varia de la etichetarea ca atare și limitarea distribuirii falsurilor profunde (excluzându-le din algoritmi de recomandare), până la ștergerea acestora. Promovarea competenței mediatică trebuie să fie, de asemenea, o prioritate pentru toți cetățenii, indiferent de vârstă. Doar înțelegând corect modul de creare și distribuire a falsurilor profunde, cetățenii vor putea recunoaște dezinformarea și vor evita să fie induși în eroare.

4.4 Responsabilitatea fiecăruia: gândirea critică și competența mediatică

Gândirea critică și competența mediatică stau la baza unei abordări diferențiate față de fenomenul dezinformării. Cu siguranță este imposibil și probabil indezirabil să rugăm fiecare persoană în parte să pună la îndoială tot ce vede.

Pe de altă parte, oamenii ar putea, mai mult ca oricând, să fie sfătuiți să consume cu precauție conținutul publicat pe internet. Cel mai simplu lucru pe care îl poate face oricine atunci când vede o imagine, un video sau un text care îi pare suspect este să caute pe Google. Deseori, conținutul manipulat va fi demascată foarte repede cu o simplă căutare, întrucât detaliile manipulării circulă la fel de repede ca și conținutul în sine.

²⁴ <https://sites.google.com/site/motormiscellany/motoring/law-and-the-motorist/locomotive-act-1865/>

Acest lucru este deosebit de important mai ales pentru utilizatorii care doresc să partajeze conținutul „apreciindu-l” sau comentându-l. La fel, putem să urmărim mai atent dacă clipirea ochilor, expresiile faciale sau vorbirea într-un video nu par firești, dacă anumite părți ale unei imagini sunt neclare sau dacă obiectele nu par la locul lor.

Aceste indicii, însă, vor dispărea rapid pe măsură ce tehnologia falsurilor profunde avansează. În viitor, teoretic ar putea să existe add-on-uri pentru browsere care să identifice automat conținutul manipulat și să-i înștiințeze pe utilizatori, asemeni unui program de blocare a reclamelor. Pentru aceasta, însă, trebuie să cunoaștem în primul rând despre posibilitatea de manipulare a conținutului.

Pentru a îmbunătăți acest tip de cunoștințe în rândul populației sale, Finlanda, țara care a fost clasată pe cea mai înaltă poziție în cadrul unui studiu de măsurare a rezistenței la dezinformare,²⁵ le oferă tuturor cetățenilor săi posibilități educaționale în acest domeniu – de la vârsta timpurie până la cea de pensionare.



²⁵ <https://osis.bg/wp-content/uploads/2019/11/MediaLiteracyIndex2019-ENG.pdf>

5.0

CE URMEAZĂ?

La momentul de față încă nu este posibil să prezicem cu exactitate consecințele concrete pe care falsurile profunde le vor avea asupra politicii și societății, însă aceasta nu înseamnă că nu trebuie să acționăm. După cum am discutat și mai sus, nici imaginile video false, și nici dezinformarea nu sunt un fenomen nou ca atare. Nouă este ușurința tot mai mare cu care un astfel de conținut poate fi creat, calitatea sa tot mai bună și capacitatea sa de distribuire.

Alegerile prezidențiale din Statele Unite, care vor avea loc în toamna anului 2020, se vor dovedi, fără îndoială, un bun test de turnesol. Recomandarea, totuși, este să nu susținem o abordare de tipul „să așteptăm și să vedem ce se întâmplă”.

Cercetătorii, companiile din domeniul tehnologiei, jurnaliștii, guvernele și înșiși utilizatorii trebuie să depună toate eforturile pentru a neutraliza impactul negativ al conținuturilor false. Primul pas este să implementăm o reglementare explicită și contramăsuri puternice împotriva pornografiei deepfake, dat fiind faptul că este deja un fenomen pe scară largă care cauzează prejudicii semnificative victimelor sale, în mare parte femei.

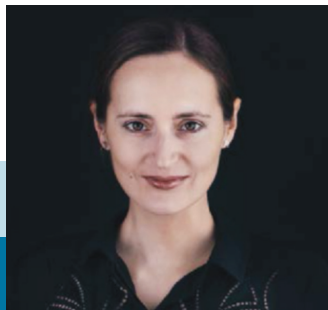


Sunt necesare și reglementări legale uniforme privind gestionarea conținutului manipulat în mass-media și pe platformele de socializare. Companii precum Facebook, Twitter, YouTube, printre altele, nu trebuie lăsate să decidă ce conținut se încadrează și ce conținut depășește domeniul de aplicare a libertății de exprimare.

Responsabilitatea față de această sarcină revine legislatorilor și democrației constituționale. Cu toate acestea, nu trebuie să cedăm tentației de a interzice complet falsurile profunde. Pe lângă riscurile sale, tehnologia deschide noi oportunități promițătoare, inclusiv în educație, film și satiră. Tehnologia în sine este neutră. Oamenii sunt cei care folosesc tehnologia fie pentru a aduce beneficii societății, fie pentru a-i face rău.

Despre autoare

Agnieszka M. Walorska



Agnieszka M. Walorska este expertă în domeniul digitalizării și Directoare Executivă la firma de gestionare și consultanță în domeniul tehnologiilor „Capco”. Ea condus mai multe proiecte de transformare și inovare pentru bănci, companii de asigurări, companii farmaceutice și companii producătoare de automobile, printre exemple.

A fondat firma de consultanță strategică digitală „CREATIVE CONSTRUCTION”, achiziționată de „Capco” în anul 2020. Este interesată în mod special de inteligența artificială și de impactul acesteia asupra interacțiunii om-mașină și, astfel, asupra modelelor de afaceri și societății.

Cu Forumul „Digital Innovation Breakfast”, ea a creat o serie de evenimente pe aceste teme, la care au luat parte speakeri proeminenți. A publicat numeroase studii și articole și participă cu regularitate ca speaker la conferințe și companii.

Agnieszka M. Walorska a contribuit la scrierea unui capitol de carte despre societatea algoritmică, în care analizează problemele etice ale inteligenței artificiale. Volumul respectiv a fost publicat de editorul științific Springer în martie 2020.

A studiat științele sociale și politice la Universitatea din Varșovia și Universitatea Humboldt din Berlin și a fost bursieră a Fundației Hertie și a Fundației Germane pentru Burse Academice.