Expunerea de motive

Libertatea Internetului nu înseamnă haos informațional - această inițiativă propune reguli clare pentru a limita răspândirea conținutului periculos, dar fără a îngrădi dreptul la exprimare, ci doar prevenind manipularea și dezinformarea la scară largă.

Internetul este si trebuie sa ramana liber, insa in conformitate cu jurisprudenta CEDO dreptul la libera exprimare, este limitat de pragul de la care exercitarea lui incalca libertatea altei persoane. Conventia Europeana a Drepturilor Omului stipuleaza expres ca exercitarea dreptului la libera exprimare comporta indatoriri si responsabilitati si este supusa unor restrictii pentru a asigura, printre altele, drepturile altora. Trebuie reglementat la nivel european acest lucru? Probabil ca da, fiindca Interentul nu are garduri, insa de undeva trebuie sa inceapa aceasta reglementare, iar Romania este cea mai indreptatita sa initieze o astfel de reglementare avand in vedere contextul anularii alegerilor prezidentiale, printre altele, pentru manipularea mediului online.

Acest proiect nu doar că nu limitează libertatea de exprimare, ci duce la crearea unui mediu sigur pentru dezbaterea reală de idei și scoate la suprafață creativitatea, conținutul original, nu pe cel toxic.

Inteligența artificială trebuie să fie un aliat, nu o amenințare - utilizată corect, poate deveni un scut împotriva conținutului nociv, ajutând la identificarea și eliminarea rapidă a dezinformării, fără a afecta fluxul liber de idei. Algoritmii de inteligență artificială sunt cei care răspândesc acum conținutul nociv, tot algoritmii îl pot opri.

Reglementare nu înseamnă cenzură, ci siguranță digital - scopul nu este interzicerea opiniilor, ci stoparea amplificării artificiale a conținutului periculos, protejând astfel consumatorii de informație din mediul online, precum și eliminarea avantajelor financiare ale celor care exploatează dezinformarea.

Aceasta initiativa legislativa nu isi propune sa inchida platformele sociale din mediul online, nici macar sa le interzica utilizatorilor sa genereze continut fals, ori sa utilizeze malitios tehnologia in contextul inteligentei arficiale. Ci isi propune sa:

- impuna furnizorilor de retele sociale limitarea propagarii continutului cu potential periculos - instiga la ura si violenta, dezinformeaza periculos pe subiecte de interes national, manipuleaza informatii cu scopul inducerii in eroare pe teme majore - asa cum este el incadrat de algoritmii de

inteligenta artificiala, **prin reducerea in cadrul algoritmului de propagare a metricii de transmitere pentru astfel de tipuri de continut;**

- sa interzica monetizarea in astfel de cazuri (astfel de continut sa nu poata fi promovat contra-cost);

- sa elimine din platforma respectiva continutul ilegal, asa cum este el definit de Regulamentul UE2022/2065, dupa un scurt interval de carantina in care este procesat si clasificat corespunzator.

Astfel, libertatea de exprimare nu este alterata, fiind chiar garantata in continuare, bucata buna a tehnologiei este accesibila in continuare doritorilor, utilizatorii sunt in continuare liberi sa se exprime absolut cum considera, doar ca, in momentul in care continutul generat este incadrat ca fiind cu potential sa instige la ura, la violenta, sa dezinformeze pe teme majore sa manipuleze informatii false cu scopul inducerii in eroare pe subiecte de interes national, acest continut sa nu se propage la mai mult de 150 de persoane. Astfel, esti liber sa spui orice, insa daca ceea ce spui tu este de natura sa prejudicieze grav populatia, mesajul tau se va transmite doar la un numar foarte mic de destinatari.

In cazul continutului ilegal, clasificare reglementată clar de Regulamentul UE2022/2065, acesta sa fie eliminat in termen de 15 minute de la publicare, interval de carantina in timpul caruia este procesat si analizat.

In momentul de fata, spatiul cibernetic este frontul pe care utilizatorul de buna credinta, fara o educatie digitala suficienta, lupta cu arme inegale in fata robotilor care utilizeaza malitios retele sociale. Propria educatie a acestui tip de utilizator nu mai este un filtru suficient de analiza, data fiind posibilitatea tehnica a multiplicarii automatizate a informatiilor false, a continutului care incita la ura, a dezinformarii pe teme majore, metoda prin care adesea utilizatorului consumator de continut i se creaza in mod fals perceptia unei alte realitati.

In fata malitiozitatii algoritmilor de inteligenta artificiala trebuie luptat tot cu algoritmi de inteligenta artificiala, nu cu mainile goale, printr-o procedura nescalabila in care pentru fiecare tip de continut considerat ilegal un utilizator trebuie sa faca cate o sesizare. Iar pentru aceasta, un guvern puternic are nevoie de instrumente prin care sa actioneze rapid, nu de proceduri nescalabile prin care lupta cu pixul pe registru pentru fiecare tip de continut ilegal.

Aditional presedintele Curtii Constitutionale a Romaniei, Marian Enache, apreciaza ca fiind sub-reglementata zona platformelor sociale, observand o serie de riscuri care deriva in societate *„Platformele social-media, fiind subreglementate, se situeaza intr-o zona volatila, care pot genera o manipulare a informatiei si o dezinformare*

masive, dificil de decriptat intr-un timp record, sau relativ scurt. Platformele de social-media, prin algorimtii inteligentei artificiale, pot juca un rol determinant in procesul de influentare a alegerilor, sau a altor tipuri de activitati, daca ele nu sunt reglementate juridic si nu indeplinesc conditia transparentei, caracteristica intrinseca a oricarei democratii. Aceste mijloace, care fac parte din urmatorul val al tehnologiei, comporta o serie de riscuri si imprevizibilitati, care, daca nu sunt controlate juridic, pot aduce mari prejudicii dificil de identificat sub aspectul generarii lor, dupa cum, prin utilizarea lor corecta, din punctul de vedere al algoritmilor, pot genera si beneficii societatii"

Da, tehnic se poate extrage bucata nociva si lasa bucata utila, trebuie doar reglementare in acest sens!

Există studii recente care explorează utilizarea inteligenței artificiale (IA) pentru detectarea conținutului periculos în rețelele sociale. Un exemplu notabil este un studiu publicat în decembrie 2023, care a evaluat eficacitatea modelelor lingvistice de mari dimensiuni (LLM) în identificarea amenințărilor publice postate online. Cercetătorii au dezvoltat instrumente personalizate pentru a colecta date de pe o comunitate online populară din Coreea, incluzând 500 de exemple non-amenințătoare și 20 de amenințări. Modelele LLM, precum GPT-3.5, GPT-4 și PaLM, au fost utilizate pentru a clasifica postările ca "amenințare" sau "sigur". Analiza statistică a arătat că toate modelele au demonstrat o acuratețe ridicată, GPT-4 obținând o acuratețe de 97,9% pentru non-amenințări și 100% pentru amenințări. Aceste rezultate sugerează că LLM-urile pot îmbunătăți moderarea conținutului la scară largă pentru a ajuta la atenuarea riscurilor online emergente.

Inteligența artificială (IA) este utilizată din ce în ce mai mult pentru a identifica și gestiona conținutul periculos de pe platformele de social media. Diverse studii au explorat capacitățile IA de a detecta materiale dăunătoare, inclusiv discursul instigator la ură, dezinformarea și conținutul legat de autovătămare.

In mai 2024, studiul autorilor Sylvia Worlali Azumah, Nelly Elsayed, Zag ElSayed, Murat Ozer, Amanda La Guardia, denumit **Deep Learning Approaches for Detecting Adversarial Cyberbullying and Hate Speech in Social Networks** s-a concentrat pe detectarea conținutului adversarial legat de cyberbullying și discursul instigator la ură în rețelele sociale. Utilizând o abordare bazată pe învățare profundă cu un algoritm de corecție, cercetătorii au obținut rezultate semnificative. Un model LSTM cu 100 de epoci a demonstrat o performanță remarcabilă, atingând o acuratețe de 87,57%, precizie de 88,73%, recall de 87,57%, scor F1 de 88,15% și un scor AUC-ROC de 91%. Performanța modelului LSTM a depășit studiile anterioare, evidențiind potențialul abordărilor de învățare profundă în detectarea și atenuarea conținutului dăunător în mediile online.

Inca din anul 2020 au aparut studii - Understanding and Detecting Dangerous Speech in Social Media - publicat de Ali Alshehri, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed **care** a dezvoltat un set de date și modele pentru a detecta discursul periculos, obținând un scor macro F1 de 59,60%, depășind semnificativ metodele de bază, sau cercetarea Detecting Radical Text over Online Media using Deep Learning a Armaan Kaur, Jaspal Kaur Saini, Divya Bansal care a folosit o rețea neuronală bazată pe Long Short-Term Memory (LSTM) pentru a detecta conținut radical în mediul online. Modelul a atins o precizie de 85,9% în clasificarea conținutului în categoriile „radical", „non-radical" și „irelevant"

Educatia este factorul esential care trebuie dezvoltat in paralel cu aceasta reglementare, fiindca spiritul critic asociat unui bagaj educational consistent este cel mai eficient filtru pentru identificarea informatiilor false, insa adesea, oportunitatile tehnologice aparute odata cu dezvoltarea mai rapida a tehologiei, sunt cu un pas in fata gradului prin care educatia reuseste sa fie unicul filtru.

Inițiator:

Radu-Dinel Miruță - deputat

Lista sustinatorilor propunerii legislative privind limitarea propagării prin platforme online foarte mari, a conținutului ilegal, care incită la ură, sau care pe teme majore de interes național, manipulează prin utilizarea malitioasa a tehnologiei, cu scopul inducerii în eroare a opiniei publice

Nr. crt	Nume si prenume	Grup parlamentar	Semnatura
1.	Durută Radu - Dinel	USR	
2.	Paraschivescu Ovidiu Romulus	USR	
3	GHEORGHIU ANDREI	USR	
4.	RIGHAN CIPRIAN	USR	
5	VASILIU DUMITRU	USR	
6	ALECSANDRU MARIUS NICOLAE	USR	
7.	ATANASIU CORINA	USR	
8	GIMA GEORGE	USR	
9	STOICA ALIN - BOGDAN	USR	
10	DIANA BUZOIANU	USR	
11	DIMITRIU ALEXANDRU	USR	
12	GIURGIU ADRIAN	USR	
13	Diana Stoica	USR	
14	Octavian Chiua	PNL	
15	ROMAN FLORIN	PNL	
16	TANASĂ ȘTEFAN	USR	
17	ECHERT ADRIAN	USR	
18	COZMA ADRIAN	PNL	
19	BOTEZ MIHAI-CĂTĂLIN	USR	

Lista sustinatorilor propunerii legislative privind limitarea propagării prin platforme online foarte mari, a conținutului ilegal, care incită la ură, sau care pe teme majore de interes național, manipulează prin utilizarea malitloasa a tehnologiei, cu scopul inducerii în eroare a opiniei publice

Nr. crt	Nume si prenume	Grup parlamentar	Semnatura
20	ȘTEFAN-ILIESCU LORINȚ	USR	
21	RUS VASILE CIPRIAN	USR	
22	SILVIU VEXLER	MINO.	

Parlamentul României

Camera Deputaților **Senatul**

LEGE

privind limitarea propagării prin platforme online foarte mari, a conținutului ilegal, care incită la ură, sau care pe teme majore de interes național, manipulează prin utilizarea malitioasa a tehnologiei, cu scopul inducerii în eroare a opiniei publice

Parlamentul României adoptă prezenta lege:

Art. 1 Prezenta lege reglementează propagarea prin platforme online foarte mari a oricărui conținut ilegal, sau care incita la ura, xenofobie, antisemitism, sau care, pe teme majore de interes național, manipulează prin utilizarea malitioasa a tehnologiei, cu scopul inducerii în eroare a opiniei publice.

Art. 2 În înțelesul prezentei legi termenii și expresiile de mai jos au următoarele semnificații:

a) platforma online foarte mare - are intelesul art. 33 din Regulamentul (UE) 2022/2065 al Parlamentului European și al Consiliului din 19 octombrie 2022 privind o piață unică pentru serviciile digitale și de modificare a Directivei 2000/31/CE (Regulamentul privind serviciile digitale)

b) rețele sociale - serviciu al unei platforme online din categoria celor definite la lit. a) care permite utilizatorilor sa creeze si/sau sa distribuie conținut către alți utilizatori;

c) furnizor de rețele sociale - o entitate care oferă și operează o platformă online din categoria celor prevăzute la lit. a) ce permite utilizatorilor să creeze profiluri, să interacționeze și să distribuie conținut digital;

d) conținut ilegal - are înțelesul art. 3, litera (h) din Regulamentul (UE) 2022/2065 al Parlamentului European și al Consiliului din 19 octombrie 2022;

e) continut cu potential periculos - are intelesul oricarui tip de continut care instiga la ura si violenta, dezinformeaza periculos, manipuleaza informatii, induce in eroare pe teme majore de interes national;

f) utilizarea malițioasă a tehnologiei - acțiunea intenționată de inducere în eroare, cu scopul de a crea dezinformare și/sau haos informațional prin manipularea comportamentului uman și/sau exploatarea vulnerabilităților cu privire la subiecte de interes major;

g) sistem de inteligență artificială - are intelesul art. 3, alin. 1 din Regulamentul UE 2024/1689 al Parlamentului European și al Consiliului din 13 iunie 2024.

Art. 3. Pentru conturile utilizatorilor cu activitate sau locatie declarate in Romania, furnizorii de retele sociale au obligatia sa isi adapteze algoritmii de propagare a continutului astfel incat continutul prevazut de art. 2, lit. e) sa nu fie transmis prin intermediul platformei online către un numar mai mare de 150 de utilizatori.

Art. 4 Pentru tipul de continut de la art. 2, lit. e) asociat conturilor utilizatorilor cu activitate sau locatie declarate in Romania, furnizorilor de retele sociale le este interzis sa intreprinda orice actiune promovare.

Art. 5 Pentru conturile utilizatorilor cu activitate sau locatie declarate in Romania, furnizorii de retele sociale au obligatia sa isi adapteze algoritmii de propagare a continutului astfel incat continutul prevazut la art. 2, lit. d) sa fie eliminat după 15 minute de la publicare, interval de carantina in care algoritmii platformei in cauza analizeaza si clasifica tipul de continut.

Art. 6 In situatia in care dupa 30 de zile de la intrarea in vigoare a prezentei legi, procentul sesizarilor cu referire la situatiile prevezute de art. 2, lit. d) si e), verificate individual de catre institutiile statului cu responsabilitati in domeniu si confirmate ca fiind indreptatite este mai mare de 30% din totalul acestora, furnizorul de retele sociale in cauza va fi sanctionat cu amenda contraventionala in valoare de 1% din cifra de afaceri.

Aceasta lege a fost adoptata de Parlamentul Romaniei, cu respectarea prevederilor art. 75 si ale art. 76 alin (2) din Constitutia Romaniei republicata.

Presedintele Camerei Deputatilor,
Serban Ciprian-Constantin

Presedintele Senatului,
Ilie-Gavril Bolojan