

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 și trebuie să rămână liber, însă în conformitate cu jurisprudența CEDO dreptul la liberă exprimare, este limitat de pragul de la care exercitarea lui încalcă libertatea altei persoane. Convenția Europeană a Drepturilor Omului stipulează expres ca exercitarea dreptului la liberă exprimare comportă îndatoriri și responsabilități și este supusă unor restricții pentru a asigura, printre altele, drepturile altora. Trebuie reglementat la nivel european acest lucru? Probabil că da, fiindcă Internetul nu are garduri, însă de undeva trebuie să înceapă această reglementare, iar România este cea mai îndreptățită să inițieze o astfel de reglementare având în vedere contextul anularii alegerilor prezidențiale, printre altele, pentru manipularea mediului online.

Acest proiect nu doar că nu limitează libertatea de exprimare, ci duce la crearea unui mediu sigur pentru dezbateră 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.

Această inițiativă legislativă nu își propune să închidă platformele sociale din mediul online, nici măcar să le interzică utilizatorilor să genereze conținut fals, ori să utilizeze malitios tehnologia în contextul inteligenței artificiale. Ci își propune să:

- impuna furnizorilor de rețele sociale limitarea propagării conținutului cu potențial periculos - instiga la ură și violență, dezinformează periculos pe subiecte de interes național, manipulează informații cu scopul inducerii în eroare pe teme majore - așa cum este el încadrat 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.

Adicional 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 algoritmi 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 bucată nocivă și lăsa bucată utilă, trebuie doar reglementare în 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.

În 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.

Încă din anul 2020 au apărut 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 „irrelevant”

Educația este factorul esențial care trebuie dezvoltat în paralel cu această reglementare, fiindcă spiritul critic asociat unui bagaj educațional consistent este cel mai eficient filtru pentru identificarea informațiilor false, însă adesea, oportunitățile tehnologice aparute odată cu dezvoltarea mai rapidă a tehnologiei, sunt cu un pas în fața gradului prin care educația reușește să fie unicul filtru.

Inițiator:

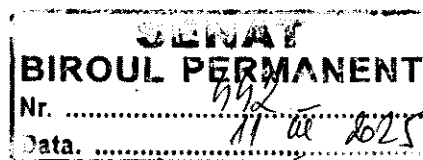
Radu-Dinel Miruță - deputat

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 malitioasă a tehnologiei, cu scopul inducerii în eroare a opiniei publice

Nr. crt	Nume si prenume	Grup parlamentar	Semnatura
1.	Ilirică Radu - Dumez	USR	
2.	Poraschivescu Ovidiu Romulus	USR	
3	GHEORGHIU ANDREI	USR	
4.	RIGHAN CIPRIAN	USR	
5	VADUVA MIRON	USR	
6	ALEXANDRU MARIUS NICOLAE	USR	
7.	ATANASIU CORINA	USR	
8	GIMA GORGO	USR	
9	STOICA ALIN - BOGDAN	USR	
10	DIANA BUZOIANU	USR	
11	Dimitrie ALEXANDRU	USR	
12	GIURGIU ADRIAN	USR	
13	Diana Spice	USR	
14	Octavian Cîrva	PNL	
15	ROMAN FLORENTIN	PNL	
16	TANASĂ ȘTEFAN	USR	
17	ECHERȚ ADRIAN	USR	
18	COZMA ADRIAN	PNL	
19	BOTEZ MIHAIL - 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 malitioasă a tehnologiei, cu scopul inducerii în eroare a opiniei publice

[illegible]



PARLAMENTUL ROMÂNIEI
CAMERA DEPUTAȚILOR
Grupul Parlamentar al USR

Către: Biroul permanent al Senatului

Subsemnatul **Ungureanu Emanuel-Dumitru**, ales deputat în circumscripția electorală nr.35 SUCEAVA, prin prezenta vă rog să-mi aprobați solicitarea de a deveni co-inițiator al *“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 malitioasă a tehnologiei, cu scopul inducerii în eroare a opiniei publice”* înregistrată la Senat în data de 05.03.2025 cu nr **B61/2025**.

Data:
11.03.2025

Semnătura:
Ungureanu Emanuel-Dumitru