**Ius et Societas**
ISSN 2587-3652 (print) | ISSN 2587-3660 (online) Vol. I(1) | 2026

DOI: https://doi.org/10.52326/ies.2026.1.08
CZU: 004.8:316.77:342.727
**Corresponding author*

OPEN ACCESS CC BY

# ALGORITHMIC CONTENT MODERATION AND FREEDOM OF EXPRESSION IN THE AGE OF ARTIFICIAL INTELLIGENCE

## MODERAREA ALGORITMICĂ A CONȚINUTULUI ȘI LIBERTATEA DE EXPRIMARE ÎN ERA INTELIGENȚEI ARTIFICIALE

**Dina BARCARI***
*Universitatea Tehnică a Moldovei. Republica Moldova*
ROR: *https://ror.org/02b82hk77*
*E-mail: dina.barcari@ssu.utm.md*
ORCID: 0000-0002-5764-6617

**Abstract:** *This study examines the relationship between algorithmic content moderation and freedom of expression in the age of artificial intelligence, highlighting the balance between protecting users from harmful content and safeguarding legitimate speech. Its main objective is to analyse how algorithmic systems shape communication practices, individual autonomy, identity formation, intergroup relations and the distribution of visibility in the digital public sphere.*
*The research is predominantly theoretical and interdisciplinary, drawing on conceptual and interpretive analysis from the perspectives of anthropology and the philosophy of technology.*
*The findings show that algorithms play an active role in structuring communication, selecting information and defining the norms of acceptable discourse. Content personalisation fosters filter bubbles, fragments the public sphere and reduces exposure to diverse perspectives. At the same time, automated classification may both allow harmful content to circulate and restrict legitimate expression. The study underlines the importance of a balanced moderation model that combines technological efficiency with human oversight.*

**Keywords:** *algorithmic moderation, artificial intelligence, freedom of expression, individual autonomy, algorithmic identity, cultural pluralism*

**Abstract:** *În această cercetare este examinată relația dintre moderarea algoritmică a conținutului și libertatea de exprimare în era inteligenței artificiale, prin care se evidențiază echilibrul dintre protejarea utilizatorilor împotriva conținutului vătămător și respectarea libertății discursului legitim. Obiectivul principal al lucrării constă în analiza modului în care sistemele algoritmice influențează practicile de comunicare, autonomia individuală, construirea identității, relațiile dintre grupuri și distribuirea*

*vizibilității în spațiul public digital.*
*Metodologia cercetării are un caracter predominant teoretic și interdisciplinar, bazându-se pe analiza conceptuală și interpretativă din perspectivele antropologiei și filosofiei tehnologiei.*
*Rezultatele analizei demonstrează că algoritmii participă activ la organizarea comunicării, selectarea informației și definirea normelor discursului acceptabil. Personalizarea conținutului favorizează formarea bulelor informaționale, fragmentarea spațiului public și reducerea contactului cu perspective diverse. Totodată, clasificarea automată permite circulația conținutului vătămător și restricționează exprimările legitime. Cercetarea evidențiază importanța unui model echilibrat de moderare prin care se îmbină eficiența tehnologică cu supravegherea umană.*

***Cuvinte-cheie:*** *moderare algoritmică, inteligență artificială, libertate de exprimare, autonomie individuală, identitate algoritmică, pluralism cultural.*

## 1. Introducere

Extinderea inteligenței artificiale în spațiul comunicării digitale transformă profund modul în care oamenii produc, distribuie, interpretează și contestă informația. Platformele digitale au devenit infrastructuri tehnice pentru schimbul de mesaje, imagini, opinii, dar și spații sociale în care se construiesc identități, comunități, forme de solidaritate și conflicte simbolice. În acest context, moderarea conținutului reprezintă una dintre cele mai importante forme de intervenție asupra comunicării contemporane. Dacă, în primele etape ale internetului, moderarea era realizată în mare măsură de utilizatori sau de echipe umane specializate, dezvoltarea inteligenței artificiale a făcut posibilă evaluarea și filtrarea automată a unor cantități uriașe de conținut. Algoritmii identifică texte, imagini, videoclipuri și comportamente considerate problematice, restricționează accesul la anumite materiale, suspendă conturi sau reduc vizibilitatea unor mesaje. Această transformare promite o mai mare eficiență în combaterea abuzului, a discursului instigator la ură, a dezinformării și a altor forme de conținut vătămător. În același timp, ea ridică întrebări fundamentale despre libertatea de exprimare, autonomie, responsabilitate, putere și natura însăși a spațiului public digital.

Problema centrală a prezentei cercetări pornește de la o tensiune aparent paradoxală: în ce măsură moderarea algoritmică protejează utilizatorii fără ca protecția să se transforme într-o formă de control al exprimării? Întrebarea devine cu atât mai complexă cu cât algoritmii nu sunt observatori neutri ai comunicării umane. Ei funcționează

pe baza unor categorii, reguli, seturi de date și criterii de clasificare construite de oameni și instituții. Prin urmare, decizia algoritmică de a considera un conținut acceptabil, ofensator, periculos sau interzis implică concepții despre normalitate, moralitate, risc și limitele discursului legitim. (Thach, H., Mayworm, S., Delmonaco, D., & Haimson, O., 2024).

Libertatea de exprimare ocupă un loc central în această dezbatere deoarece constituie una dintre condițiile fundamentale ale vieții politice și sociale moderne. În tradiția filosofică, libertatea de formulare și comunicare a opiniilor este asociată cu autonomia persoanei și cu posibilitatea participării la viața publică. De la apărarea libertății intelectuale la concepția despre „piața ideilor", numeroase teorii filosofice au considerat exprimarea liberă drept un mecanism indispensabil pentru contestarea autorității, corectarea erorilor și dezvoltarea cunoașterii. Totuși, libertatea de exprimare nu a fost niciodată absolută. Societățile au stabilit limite în cazul amenințărilor, defăimării, incitării la violență sau al altor forme de comunicare considerate vătămătoare. Ceea ce se schimbă în mediul digital este scara, viteza și mecanismul prin care aceste limite sunt aplicate.

Moderarea algoritmică introduce o formă particulară de mediere între individ și spațiul public. Atunci când o platformă elimină automat o postare sau reduce vizibilitatea unui anumit tip de conținut, utilizatorul nu interacționează doar cu o regulă explicită, ci cu un sistem de clasificare care interpretează discursul său. Această situație este deosebit de problematică deoarece limbajul uman este profund dependent de context. Ironia, satira, metafora, citatul, referința culturală, umorul sau critica politică sunt dificil de diferențiat automat de conținutul ofensator ori periculos. Un algoritm identifică anumite cuvinte sau tipare statistice, însă semnificația unui enunț depinde adesea de cine îl produce, cui îi este adresat, în ce situație este formulat și ce intenție comunicativă are. Astfel, ceea ce apare algoritmului drept o unitate de text reprezintă, pentru un subiect uman, o experiență socială complexă. Riscul devine și mai semnificativ atunci când algoritmii funcționează opac. Utilizatorul uneori nu știe de ce un anumit conținut a fost eliminat, de ce contul său a fost sancționat sau de ce mesajul său nu mai ajunge la public. O astfel de opacitate creează un dezechilibru între platformă și utilizator: prima dispune de infrastruc-

tura tehnică și de capacitatea de interpretare a datelor, în timp ce a doua are acces limitat la explicații și posibilități de contestare.

Miza cercetării depășește problema tehnică a moderării conținutului. Punem în discuție însăși natura spațiului public într-o societate în care o parte tot mai importantă a comunicării este intermediată de sisteme algoritmice. Dacă în societățile moderne libertatea de exprimare a fost gândită în raport cu statul, presa și instituțiile publice, în era inteligenței artificiale ea trebuie regândită și în raport cu platformele digitale, modelele de învățare automată și infrastructurile private care organizează circulația informației. Algoritmii devin astfel actori indirecți ai vieții sociale și politice, iar întrebarea privind libertatea se transformă într-o întrebare despre cine are puterea de a decide ce poate fi spus, auzit, văzut și transmis mai departe (Gorwa et al., 2020).

Prin urmare, studierea moderării algoritmice reprezintă o oportunitate de reexaminare a celor mai vechi preocupări sociale și anume complexitatea limbajului uman, diversitatea culturală și dreptul la contestare.

## 2. Stadiul actual al cercetării și scopul acesteia

Cercetarea moderării algoritmice și a libertății de exprimare reprezintă un domeniu interdisciplinar, situat la intersecția dintre antropologie, filosofie, comunicarea digitală și etica inteligenței artificiale. Dezvoltarea platformelor digitale și utilizarea algoritmilor pentru identificarea, clasificarea și eliminarea conținutului problematic au transformat modul în care informația este produsă, distribuită și receptată. Deși aceste tehnologii contribuie la combaterea discursului instigator la ură, a hărțuirii și a dezinformării, utilizarea lor ridică întrebări importante privind libertatea de exprimare, autonomia individuală și diversitatea culturală.

În cadrul cercetărilor realizate anterior asupra acestei problematici, s-a analizat relația dintre om și tehnologie, rolul algoritmilor în organizarea comunicării digitale și implicațiile moderării automate asupra vieții sociale. S-a urmărit înțelegerea în ce măsură sistemele de inteligență artificială contribuie la protejarea utilizatorilor și în ce condiții intervenția lor devine o formă de control al discursului. Considerăm că această problemă trebuie să fie abordată din perspectivă antropologică, deoarece algoritmii influențează și modul în care indi-

vizii își exprimă identitatea, își construiesc opiniile și interacționează cu ceilalți.

Analiza pornește de la ideea că platformele digitale au devenit medii în care se organizează vizibilitatea informației și condițiile participării la spațiul public. Prin recomandări, clasificări și restricții, algoritmii influențează mesajele care ajung la public și perspectivele cu care utilizatorii intră în contact. În acest context, am acordat o atenție deosebită relației dintre moderarea algoritmică și autonomia umană, evidențiind faptul că libertatea publicării unui mesaj nu garantează automat posibilitatea ca acesta să fie văzut și înțeles de ceilalți.

Prin cercetarea realizată, ne propunem să susținem necesitatea unui model echilibrat de moderare, care să combine eficiența tehnologiei cu evaluarea umană, transparența și responsabilitatea. În viziunea noastră protecția utilizatorilor trebuie să fie compatibilă cu respectarea diversității discursului și cu dreptul individului de a-și exprima opiniile.

## 3. Metode și materiale

În cadrul articolului, a fost utilizată o abordare interdisciplinară, bazată pe metode de cercetare teoretică și conceptuală, pentru analiza implicațiilor antropologice și filosofice ale moderării algoritmice asupra libertății de exprimare. Alegerea acestor metode a fost determinată de caracterul complex al temei, care presupune examinarea atât a funcționării tehnologiilor digitale, cât și a efectelor acestora asupra individului și societății.

Metoda analizei teoretice a fost aplicată pentru examinarea conceptelor fundamentale ale cercetării, precum inteligența artificială, moderarea algoritmică, libertatea de exprimare, autonomia, identitatea și alteritatea. Prin această metodă, a fost evidențiată relațiile dintre conceptele analizate și s-a urmărit explicarea modului în care sistemele algoritmice influențează comunicarea și participarea indivizilor la spațiul public digital.

Metoda analizei conceptuale și filosofice a fost utilizată pentru examinarea implicațiilor moderării automate asupra libertății individuale și responsabilității umane. A fost analizată tensiunea dintre necesitatea protejării utilizatorilor împotriva conținutului vătămător și riscul restricționării exprimării legitime. Această metodă a permis

să formulăm întrebări privind legitimitatea deciziilor algoritmice, autonomia individului și posibilitatea contestării intervențiilor automate.

Abordarea antropologică a fost aplicată pentru interpretarea moderării algoritmice ca fenomen social și cultural. S-a urmărit evidențierea modului în care limbajul, identitatea și apartenența la comunități influențează semnificația mesajelor digitale.

Metoda analizei critice a fost utilizată pentru examinarea consecințelor sociale și etice ale utilizării inteligenței artificiale în moderarea conținutului. S-a evidențiat riscurile asociate erorilor algoritmice, opacității deciziilor și distribuirii inegale a vizibilității în mediul digital (Gorwa et al., 2020). Totodată, s-a analizat rolul platformelor în stabilirea normelor de comunicare și în delimitarea conținutului considerat acceptabil.

Metoda sintezei a permis integrarea perspectivelor filosofice și antropologice într-o interpretare unitară a fenomenului cercetat. Prin corelarea conceptelor analizate, s-a urmărit identificarea principiilor necesare unei moderări responsabile, precum transparența, responsabilitatea umană, posibilitatea contestării deciziilor și respectarea diversității culturale.

Articolul urmărește să demonstreze că moderarea algoritmică reprezintă atât o dimensiune tehnologică, cât și una socială, culturală și morală. Abordarea adoptată susține necesitatea unui echilibru între protecția utilizatorilor și respectarea libertății de exprimare, evidențiind importanța implicării judecății umane în situațiile în care interpretarea contextului este esențială.

## 4. Rezultate și Discuții

### 4.1 Omul și tehnologia: transformarea spațiului comunicării

Relația dintre om și tehnologie reprezintă una dintre constantele fundamentale ale dezvoltării societăților umane. De la primele instrumente utilizate pentru procurarea hranei și protejarea corpului până la infrastructurile digitale contemporane, omul a creat tehnologii care, la rândul lor, au modificat condițiile în care acesta trăiește, comunică și gândește. În acest sens, tehnologia reprezintă o componentă integrată în procesele prin care cultura este produsă și transmisă.

Comunicarea constituie unul dintre domeniile culturii în care această relație devine deosebit de vizibilă. Limbajul însuși este con-

siderat o tehnologie socială prin care experiențele individuale sunt transformate în semnificații accesibile celorlalți. Apariția scrisului a permis conservarea și transmiterea informației dincolo de limitele memoriei individuale și ale comunicării orale. Tiparul a extins radical posibilitatea distribuirii ideilor, contribuind la apariția unor noi forme de public și la accelerarea circulației cunoașterii. Telegraful, telefonul, radioul și televiziunea au modificat ulterior raporturile dintre distanță, timp și comunicare. Pe când, internetul a produs o transformare și mai profundă, deoarece a făcut posibilă comunicarea simultană, globală și interactivă a unui număr enorm de persoane. În prezent, inteligența artificială adaugă o nouă dimensiune acestei evoluții prin care comunicarea este interpretată, clasificată, filtrată și chiar generată de acestea.

Transformarea în cauză impune reconsiderarea modului în care definim spațiul comunicării. În modelul tradițional al comunicării, individul era conceput ca subiect care produce un mesaj și îl transmite unui alt subiect (emițător-receptor). Chiar dacă între cei doi existau instituții și tehnologii intermediare, relația păstra într-o anumită măsură caracterul unei interacțiuni între persoane. În mediul digital, această structură este mult mai complexă. Mesajul circulă prin platforme care colectează date, analizează comportamente, estimează preferințe și stabilesc, prin sisteme algoritmice, cui și în ce condiții îi va fi afișat conținutul. Comunicarea devine astfel o relație între oameni și infrastructuri tehnologice care participă la organizarea schimbului simbolic.

Din această perspectivă, platforma digitală stabilește condițiile în care comunicarea are loc. Interfața, algoritmii de recomandare, sistemele de moderare, regulile comunității sunt adesea rezultatul interacțiunii dintre preferințele utilizatorului și arhitectura tehnologică a platformei (de exemplu, alegerea unui videoclip, a unei postări sau a unui subiect de discuție). Astfel, libertatea de comunicare în spațiul digital este întotdeauna exercitată într-un mediu deja structurat de anumite reguli tehnice și economice (Gorwa et al., 2020; Wright, 2022).

Una dintre cele mai importante schimbări produse de tehnologie este, prin urmare, transformarea raportului dintre vizibilitate și putere. În spațiul social tradițional, accesul la un public larg era limitat de existența unor instituții precum presa, editurile, radioul sau televi-

ziunea. Internetul a redus anumite bariere de acces și a permis indivizilor să publice și să distribuie conținut fără intermediari tradiționali. Această democratizare aparentă a comunicării a fost însă însoțită de apariția unor noi forme de intermediere. Platformele digitale au devenit infrastructurile prin care conținutul este ordonat și distribuit.

Aici apare importanța algoritmilor. Algoritmul reprezintă, într-un sens tehnic, o procedură de procesare a informației, însă consecințele sale sociale depășesc cu mult dimensiunea matematică. Atunci când un algoritm decide ce conținut apare în partea superioară a unui flux, ce material este recomandat unui utilizator sau ce postare este eliminată, el participă la structurarea experienței sociale. Utilizatorul nu întâlnește pur și simplu „internetul", ci o versiune a acestuia organizată de sisteme automate. Realitatea digitală devine astfel una mediată algoritmic (Zeng & Kaye, 2022).

Relația dintre om și algoritm este una dialectică. Pe de o parte, omul creează algoritmul, stabilește obiectivele sale și îi furnizează datele necesare funcționării. Pe de altă parte, algoritmul transformă comportamentul uman prin recomandări, clasificări și recompense. Se formează astfel un circuit de feedback în care utilizatorii generează date, datele contribuie la perfecționarea sistemelor, iar sistemele modifică ulterior comportamentul utilizatorilor. Această relație este importantă pentru înțelegerea inteligenței artificiale deoarece arată că aceasta nu funcționează într-un vid social. Sistemele algoritmice sunt integrate într-o relație continuă cu practicile, valorile și conflictele societății.

În cazul moderării conținutului, această relație devine deosebit de sensibilă; sistemele identifică discursul problematic. Dar această clasificare implică o concepție despre normă. Ce este ofensator? Ce este amenințător? Ce este satiră? Ce reprezintă dezinformarea? Ce diferențiază critica politică de incitare? Aceste întrebări nu au întotdeauna răspunsuri universale și obiective. Ele sunt influențate de contextul cultural, de valorile societății și de relațiile de putere. Când aceste categorii sunt transformate în criterii computaționale, complexitatea lor socială este tradusă în reguli care trebuie aplicate la scară largă. O regulă culturală este transformată într-un parametru tehnic, iar un principiu moral devine o categorie de clasificare. Această materializare conferă normei o aparență de obiectivitate. Dacă o persoană decide că un mesaj este ofensator, direct se contestează judecata ace-

lei persoane. Dacă un sistem automat ia aceeași decizie, aceasta pare rezultatul unei proceduri impersonale și obiective. În realitate, însă, algoritmul reflectă alegerile celor care l-au proiectat, ale celor care au definit categoriile și ale datelor pe care a fost antrenat.

Aici se află una dintre cele mai importante probleme filosofice ale erei inteligenței artificiale: delegarea judecății umane către sisteme automate. Judecata presupune, în mod tradițional, interpretare, deliberare și asumarea responsabilității. Algoritmizarea acesteia produce impresia că decizia este mai rapidă și mai obiectivă, însă ridică întrebarea dacă eficiența înlocuiește deliberarea. În cazul moderării, problema este cu atât mai sensibilă cu cât obiectul judecății îl constituie expresia umană, adică una dintre principalele forme prin care individul participă la viața socială.

O societate care deleagă o parte semnificativă a controlului discursului către sisteme automate riscă, astfel, să transforme libertatea de exprimare într-o libertate condiționată de clasificarea algoritmică. Aceasta nu înseamnă că orice intervenție algoritmică este ilegitimă. Dimpotrivă, utilizarea inteligenței artificiale contribuie la protejarea utilizatorilor și la reducerea unor forme de abuz care ar fi imposibil de gestionat exclusiv prin intervenție umană. În acest punct, relația dintre om și tehnologie trebuie înțeleasă ca una de co-construcție. Oamenii proiectează tehnologiile, dar tehnologiile contribuie ulterior la modelarea practicilor umane. În cazul comunicării digitale, această co-construcție este deosebit de vizibilă: omul își adaptează discursul la platformă, iar platforma își adaptează algoritmii la comportamentul utilizatorilor. Se formează astfel un spațiu comunicativ hibrid, în care granița dintre social și tehnic devine din ce în ce mai dificil de trasat (Christin, A., 2020).

Din această perspectivă, întrebarea privind viitorul comunicării depinde de întrebarea privind viitorul relației dintre om și inteligența artificială. Dacă algoritmii vor participa din ce în ce mai mult la selectarea, filtrarea și generarea discursului, atunci va fi necesară o reflecție critică asupra principiilor care trebuie să guverneze această participare. Transparența, posibilitatea de contestare, pluralismul cultural, protecția împotriva discriminării și păstrarea autonomiei individuale devin condiții esențiale ale unei comunicări digitale compatibile cu valorile democratice.

Tehnologia oferă oamenilor posibilități extraordinare de exprimare, conectare și participare, dar introduce simultan noi forme de dependență, supraveghere și control. Inteligența artificială accentuează, astfel, această ambivalență deoarece funcționează atât ca instrument de protecție, cât și ca mecanism de filtrare a discursului. Din acest motiv, analiza relației dintre om și tehnologie, în viziunea noastră, depășește întrebarea „ce poate face tehnologia?" și se concentrează asupra întrebării „ce fel de societate construim prin modul în care alegem să folosim tehnologia?"

În raport cu tema generală a moderării algoritmice și a libertății de exprimare, această întrebare dobândește o importanță fundamentală. În acest mod, analizăm tipul de relație socială pe care aceste sisteme îl produc între individ, comunitate și autoritatea care stabilește limitele discursului. Legitimitatea tehnologiei depinde de valorile pe care le încorporează, de distribuția puterii pe care o produce și de posibilitatea oamenilor de a participa la stabilirea regulilor care le organizează spațiul comunicativ.

Astfel, transformarea tehnologică a comunicării reprezintă, în esență, și o transformare antropologică. Prin tehnologiile noi oamenii devin subiecți ai unor noi forme de comunicare, identitate și relaționare. În același timp, apar forme de autoritate, iar algoritmii devin intermediari între expresia individuală și publicul căruia aceasta îi este destinată. Tocmai de aceea, problema moderării algoritmice nu este redusă la o discuție despre software, securitate sau eficiență. Ea este o problemă despre om, libertate și putere (Seaver, N., 2017).

În era inteligenței artificiale, apărarea libertății de exprimare presupune, prin urmare protejarea condițiilor sociale și tehnologice care fac posibil ca vocea individului să fie auzită. În același timp, protecția utilizatorilor presupune dezvoltarea unor mecanisme capabile să limiteze prejudiciile fără să transforme diversitatea discursului uman într-o problemă de eliminat. Între aceste două imperative se construiește noul spațiu al comunicării digitale, un spațiu în care omul și tehnologia se influențează reciproc și în care problema libertății devine inseparabilă de problema modului în care tehnologiile sunt proiectate, utilizate și guvernate.

### 4.2 Libertatea de exprimare ca formă a autonomiei umane

Libertatea de exprimare constituie una dintre formele fundamentale prin care autonomia umană se manifestă în spațiul social. Capacitatea individului prin care formulează ideile, exprimă judecăți, pune sub semnul întrebării normele existente și de comunicare a propriilor experiențe reprezintă mai mult decât exercitarea unui drept juridic. Ea este legată de însăși capacitatea persoanei de a se constitui ca subiect prin participarea la viața comunității. Din această perspectivă, libertatea de exprimare nu trebuie analizată doar ca libertate de a produce și transmite mesaje, ci ca una dintre condițiile prin care individul își afirmă autonomia, identitatea și demnitatea.

Conceptul de autonomie ocupă un loc central în filosofia modernă. În sens larg, autonomia desemnează capacitatea persoanei de a-și conduce propria existență, de a formula judecăți și de a acționa în conformitate cu motive pe care le consideră ale sale. Autonomia presupune, astfel, o anumită independență față de constrângeri externe, dar nu trebuie confundată cu izolarea individului față de societate. Omul este o ființă socială, iar formarea opiniilor, valorilor și identității sale are loc în relație cu ceilalți. Prin urmare, autonomia înseamnă posibilitatea participării în mod reflexiv la procesele prin care aceste influențe sunt acceptate, contestate sau transformate.

În acest sens, exprimarea reprezintă una dintre condițiile fundamentale ale autonomiei. Pentru a-și exercita capacitatea de judecată, individul trebuie să poată formula și comunica ceea ce gândește. O persoană căreia îi este interzis sistematic să își exprime opiniile este lipsită nu doar de posibilitatea comunicării, ci și de o parte importantă a capacității sale de a participa la definirea propriei poziții sociale. Exprimarea permite individului să transforme experiența interioară într-o realitate socială. Prin cuvinte, gesturi, imagini, simboluri și alte forme de comunicare, persoana își face cunoscută perspectiva și intră într-o relație de recunoaștere cu ceilalți (Seaver, N., 2018).

Libertatea de exprimare trebuie, prin urmare, înțeleasă ca o relație între individ și comunitate. Omul vorbește către ceilalți, în raport cu ceilalți și adesea împotriva unor poziții deja existente. Exprimarea are astfel o dimensiune dialogică. Ea presupune trei criterii: de a răspunde, de a critica și de a fi criticat. Tocmai această dimensiune face ca libertatea de exprimare să fie esențială pentru existența

unui spațiu public pluralist. În absența posibilității de a întâlni opinii diferite, autonomia riscă să fie redusă la o simplă libertate privată, lipsită de capacitatea de a participa la transformarea lumii comune.

Spațiul digital a extins considerabil posibilitățile de exprimare, oferind indivizilor acces la audiențe care, în trecut, erau dificil de atins. O persoană își expune opinia, participă la o dezbatere, creează comunități sau atrage atenția asupra unei probleme sociale fără să depindă în mod necesar de instituțiile tradiționale ale comunicării. În acest sens, mediul digital contribuie la democratizarea exprimării și la diversificarea vocilor prezente în spațiul public. Totuși, această democratizare este însoțită de noi forme de mediere. Libertatea formală de a publica nu garantează și posibilitatea efectivă de a fi auzit. În mediul digital, vizibilitatea depinde în mare măsură de infrastructurile tehnologice care organizează circulația conținutului. Algoritmii determină ce apare în fluxul unui utilizator, ce este recomandat și ce este eliminat sau redus ca vizibilitate. Prin urmare, între libertatea producerii unui mesaj și libertatea participării efective la comunicarea publică apare un nou nivel de mediere algoritmică.

Această distincție este fundamentală. Dacă libertatea de exprimare este analizată doar ca drept de a produce conținut, atunci faptul că un mesaj este încărcat pe o platformă poate părea suficient. Dar dacă exprimarea este înțeleasă ca formă de participare socială, atunci trebuie analizate și condițiile în care mesajul ajunge la un public. Un conținut care rămâne formal disponibil, dar care este sistematic ascuns, deprioritizat sau eliminat din recomandări are o capacitate redusă de a participa la dezbaterea publică. În acest caz, problema libertății nu se mai limitează la întrebarea „poți să vorbești?", ci devine „poți fi auzit?".

Această transformare modifică și sensul conceptului de autonomie care presupune și accesul la condițiile cognitive și sociale necesare pentru formarea opiniei. Individul are nevoie de informații, de posibilitatea de a întâlni perspective diferite și de capacitatea de a reflecta critic asupra acestora. Dacă mediul comunicativ este puternic personalizat și dacă algoritmii selectează informația în funcție de comportamentul anterior al utilizatorului, există riscul ca acesta să fie expus în mod repetat unor perspective similare. În asemenea condiții, autonomia afectează prin configurarea mediului informațional în care opiniile sunt formate.

Această problemă este apropiată de ceea ce numim autonomie cognitivă. Un individ autonom este capabil să reflecteze asupra motivelor pentru care gândește într-un anumit fel. În societatea digitală, această capacitate este influențată de arhitectura informațională. Recomandările automate, notificările, selecția conținutului și mecanismele de personalizare orientează atenția și creează anumite priorități cognitive. Desigur că, mediul informațional este organizat de sisteme care anticipează și modelează comportamentul utilizatorului, justifică o analiză critică a condițiilor în care autonomia este exercitată.

Din perspectivă antropologică, autonomia trebuie înțeleasă și în relație cu identitatea. Exprimarea este unul dintre principalele mecanisme prin care indivizii își construiesc și negociază identitatea socială. O persoană se definește prin poveștile pe care le spune despre sine, prin grupurile cu care se identifică, prin valorile pe care le afirmă și prin pozițiile pe care le asumă. În mediul digital, aceste procese sunt amplificate deoarece identitatea se modifică și se negociază în raport cu un public larg. Profilul digital devine astfel un spațiu de reprezentare a sinelui (Thach, H., Mayworm, S., Delmonaco, D., & Haimson, O., 2024).

Atunci când anumite forme de exprimare sunt eliminate automat, problema depășește simpla pierdere a unui mesaj. Pentru anumite persoane și comunități, discursul reprezintă un instrument de recunoaștere și afirmare identitară. O expresie, un simbol sau o practică discursivă care pare problematică din perspectiva unui sistem general de clasificare are o semnificație particulară pentru o comunitate. De aici rezultă una dintre dificultățile fundamentale ale moderării algoritmice: cum protejăm libertatea de exprimare într-un spațiu caracterizat de o diversitate enormă de experiențe și coduri culturale? E o întrebare prin care identificăm că autonomia este o parte componentă a pluralismului. O societate pluralistă presupune tocmai existența unor forme multiple de vorbire, de interpretare a lumii și de exprimare a experienței. Totuși, pluralismul acceptă comportamente care respectă libertatea Celuilalt.

În cazul moderării algoritmice sistemul trebuie să transforme judecăți normative în categorii tehnice. În cazul unui om o afirmație este interpretată de către acesta în funcție de context, ton, intenție și relația dintre interlocutori. Chiar și atunci când tehnologia devine performantă, ea întâlnește situații ambigue în care aceeași expresie

are mai multe semnificații. În consecință, există riscul ca mecanismul de protecție să producă restricții excesive.

Aici aducem în atenție *eroarea algoritmică* care este de două forme. Prima constă în eșecul eliminării unui conținut realmente vătămător. A doua ține de eliminarea unui conținut legitim. Dacă primul tip de eroare permite perpetuarea abuzului, cel de-al doilea afectează libertatea de exprimare. Cele două riscuri sunt dificil de echilibrat, deoarece reducerea unuia duce uneori la creșterea celuilalt. În viziunea noastră, o soluție etică ar fi politica moderată a protejării utilizatorilor împotriva formelor de abuz coordonată de un specialist care utilizează un instrument digital.

### 4.3 Algoritmul ca mediator și producător al realității sociale

În societatea contemporană, algoritmul a devenit una dintre infrastructurile invizibile prin care realitatea socială este organizată, interpretată și experimentată. Deși termenul „algoritm" provine din vocabularul matematic și informatic, utilizarea sa actuală depășește cu mult domeniul calculului. Algoritmii selectează informații, clasifică persoane și comportamente, recomandă conținuturi, anticipează preferințe și contribuie la stabilirea condițiilor în care indivizii interacționează. În mediul digital, algoritmul devine un mediator între individ și lumea socială. Prin intermediul său, oamenii întâlnesc o realitate deja selectată, ordonată și ierarhizată.

Înțelegerea algoritmului ca mediator presupune depășirea concepției potrivit căreia tehnologia ar reprezenta un simplu canal pasiv între utilizator și informație. Într-un model tradițional al comunicării, un individ produce un mesaj, acesta este transmis printr-un canal, iar un alt individ îl recepționează. În mediul digital, între producător și receptor există sisteme care decid ce conținut este afișat, în ce ordine, cu ce frecvență și către ce categorie de utilizatori este orientat. Astfel, algoritmul intervine în procesul de comunicare fără să fie neapărat vizibil (Jereza, R., 2021).

Această mediere transformă raportul dintre informație și cunoaștere. Într-un mediu în care cantitatea de informație disponibilă depășește cu mult capacitatea de procesare a unui individ, selecția devine inevitabilă. Nimeni nu poate vedea toate mesajele produse într-un spațiu digital global. Algoritmii devin, prin urmare, mecanisme

de orientare în acest exces informațional. Ei stabilesc ce este relevant pentru fiecare utilizator pe baza unor criterii care includ comportamentul anterior, interesele presupuse, interacțiunile sociale și probabilitatea ca un anumit conținut să genereze atenție. În acest fel, algoritmul participă la construirea mediului informațional în care individul își formează opiniile.

Din perspectivă antropologică, această transformare este semnificativă, așa cum în orice societate, oamenii utilizează instituțiile și practicile pentru a determina ce informații sunt importante și ce forme de cunoaștere sunt legitime. În societățile tradiționale, aceste funcții erau îndeplinite de familie, comunitate, instituții religioase, lideri locali sau forme de autoritate culturală. În societățile moderne, presa, școala, universitatea și instituțiile politice au un rol important în structurarea spațiului informațional. În societatea digitală, algoritmii devin actori suplimentari în această distribuție a relevanței.

Puterea algoritmică este, însă, diferită de formele tradiționale de putere. Ea nu se manifestă întotdeauna printr-o comandă explicită sau printr-o interdicție directă, ci funcționează prin ordonare, recomandare, prioritizare și personalizare. Un anumit conținut este prezentat de mai multe ori, altul poate fi împins într-o zonă mai puțin vizibilă, iar un al treilea este eliminat complet. Aceste intervenții produc efecte diferite, dar toate contribuie la configurarea experienței utilizatorului.

În acest sens, privim algoritmul ca o formă de „putere infrastructurală". Algoritmul structurează mediul în care se întâlnesc ideile și se formează comportamentul online. Spre exemplu, când un utilizator este expus constant anumitor tipuri de conținut, atunci acestea devin familiare lui. Dacă alte perspective sunt întâlnite rar, ele par marginale chiar dacă au o relevanță reală. Astfel, utilizatorul vede o anumită versiune a lumii digitale, însă nu vede neapărat criteriile prin care aceasta a fost construită. Fluxul personalizat creează impresia că anumite opinii sunt dominante, că unele subiecte sunt mai importante sau că alte evenimente sunt mai frecvente decât par în realitate. În acest sens, algoritmul produce o „realitate percepută" care este diferită de totalitatea informației existente.

Din cele expuse anterior deducem că personalizarea digitală de către algoritm afectează gândirea critică și produce opinii distorsionate. Dacă fiecare individ primește o versiune diferită a spațiului

informațional, atunci experiența comună a realității este dificil de menținut. Societatea are nevoie de anumite puncte de referință comune pentru a delibera asupra problemelor colective. Când indivizii sunt separați în medii informaționale diferite, dialogul este afectat. În acest sens punem în discuție formarea *bulelor informaționale* și fragmentarea spațiului public (Meppelink et al., 2024). Un utilizator este expus predominant unor idei compatibile cu interesele și comportamentele sale anterioare. În timp, această selecție reduce întâlnirea cu alte interese și oportunități. Din perspectiva autonomiei, situația este problematică deoarece libertatea în formarea opiniei presupune compararea și evaluarea alternativelor.

În acest sens, sistemele digitale colectează informații despre comportamentul utilizatorilor și formulează predicții despre preferințele lor. Uneori, sistemul anticipează interesele unei persoane înainte ca aceasta să le formuleze conștient. Considerăm că apare o dilemă etică despre identitate și autonomie: dacă un algoritm prezice ceea ce caută utilizatorul, cumpăra sau urmărește, în ce măsură preferințele omului sunt stabilite și produse prin interacțiunea cu mediul?

Totuși, faptul că un sistem anticipează un comportament nu înseamnă că individul este lipsit de libertate. Algoritmul trece de la rolul de observator la cel de producere în momentul când utilizatorul acționează într-un anumit mod repetabil. Mecanismul în cauză este esențial pentru înțelegerea societății digitale. Utilizatorul generează date prin activitatea sa, iar aceste date sunt utilizate pentru a construi modele predictive. Modelele produc apoi recomandări care influențează activitatea ulterioară a utilizatorului. Se formează astfel un circuit de feedback între comportamentul uman și sistemul algoritmic. Realitatea socială digitală este produsă prin această interacțiune continuă.

În cazul moderării conținutului, circuitul de feedback funcționează într-un mod similar. Utilizatorii învață ce tipuri de conținut sunt acceptate și ce tipuri sunt sancționate. Ei își modifică discursul în funcție de experiențele anterioare, iar aceste modificări generează noi date pentru sistem. În timp, se produc adaptări reciproce între algoritm și utilizator. Normele platformei devin astfel parte din comportamentul utilizatorilor, iar comportamentul acestora contribuie la stabilizarea normelor algoritmice. Această dinamică demon-

strează că algoritmul trebuie perceput ca un „actor socio-tehnic", situat la intersecția dintre oameni, instituții, date, infrastructuri și valori. În spatele fiecărui sistem algoritmic există organizații care îl dezvoltă, finanțează și administrează. Există moderatori umani, ingineri, juriști, specialiști în politici publice și utilizatori care contribuie, în mod direct sau indirect, la funcționarea sa. A vorbi despre „decizia algoritmului" fără a analiza aceste relații riscăm să înlăturăm responsabilitatea umană.

Din această perspectivă, prin antropologie demitizăm algoritmul. În loc să îl tratăm ca pe o entitate neutră și autonomă, îl analizăm prin contextul relațiilor sociale în care este produs și utilizat. Cine decide ce trebuie să detecteze algoritmul? Cine stabilește ce constituie conținut problematic? Cine este protejat de regulile sale? Cine este mai vulnerabil la erori? Cine contestă deciziile? Prin aceste întrebări mutăm analiza de la tehnologie ca obiect către tehnologie ca relație socială.

Apariția inteligenței artificiale ridică o întrebare mult mai profundă: *ce se întâmplă cu omul atunci când procesele fundamentale ale cunoașterii, comunicării, reprezentării și deciziei sunt mediate de sisteme non-umane*? În această perspectivă, problema autonomiei, a identității și a alterității devine centrală, deoarece inteligența artificială intervine nu doar asupra modului în care omul utilizează tehnologia, ci și asupra modului în care acesta se percepe pe sine și îi percepe pe ceilalți.

Așa cum omul face parte dintr-un mediu cultural, identitatea sa se formează prin interacțiune cu familia, comunitatea, instituțiile, limbajul, tradițiile și formele de comunicare disponibile într-o anumită societate. Tehnologia digitală și inteligența artificială devin, în prezent, componente ale acestui mediu și contribuie la construirea imaginii despre sine a individului. Identitatea contemporană este, într-o măsură semnificativă, și o identitate digitală. Profilurile online, fotografiile, comentariile, preferințele, istoricul de navigare, interacțiunile și comportamentele de consum alcătuiesc reprezentări parțiale ale persoanei. Aceste reprezentări sunt prelucrate algoritmic și transformate în categorii operaționale:

<<utilizator interesat de anumite produse>>

<<persoană susceptibilă de a interacționa cu un anumit tip de conținut>>

<<membru al unei anumite categorii demografice sau comportamentale>>

Astfel, individul devine obiectul unei clasificări permanente care produce așa numita „identitate algoritmică". Identitatea algoritmică nu coincide complet cu identitatea trăită, dar o influențează pe cea reală. Dacă un sistem presupune că o persoană este interesată de un anumit tip de conținut, îi va oferi mai mult din acel conținut. În timp, individul fiind expus repetitiv acelorași perspective și ajunge să creadă că acestea corespund în mod natural propriilor sale interese. Se produce astfel o relație circulară între preferințele utilizatorului și predicțiile sistemului.

Din moment ce sistemul anticipează nevoile utilizatorului, spațiul deliberării devine mai redus. În loc omul să caute, să compare și să descopere, el ajunge să primească. În loc să formuleze o preferință, ajunge să confirme una sugerată anterior. Astfel, considerăm că autonomia este erodată prin diminuarea exercițiului deliberativ. În același timp, este afectată și alteritatea. Alteritatea în antropologie desemnează raportarea la Celălalt și la ceea ce este diferit de sine. Identitatea se construiește în relație cu Alteritatea, așa cum omul se descoperă pe sine prin întâlnirea cu celălalt, prin dialog, conflict, diferență și recunoaștere. Personalizarea algoritmică însă reduce contactul cu alteritatea. Dacă sistemele sunt proiectate pentru a furniza utilizatorului conținut compatibil cu interesele sale anterioare, experiența digitală devine din ce în ce mai omogenă. Utilizatorul vede ceea ce probabil îi place, citește opinii apropiate de cele pe care le-a exprimat anterior și interacționează cu persoane care împărtășesc valori sau preferințe similare. Într-un asemenea mediu, diferența este percepută ca o perturbare a dialogului cu Celălalt.

Riscul este apariția unor spații de comunicare în care alteritatea devine invizibilă. Individul are impresia că propria perspectivă este dominantă sau evidentă, deoarece sistemul îi oferă permanent confirmări ale acesteia. Fenomenul este relevant atât pentru viața politică, cât și pentru relațiile culturale și sociale. O societate democratică presupune existența pluralității, iar pluralitatea este posibilitatea reală ca opiniile diferite să se întâlnească în spațiul public.

Din acest punct de vedere, algoritmul devine un mediator al alterității. El decide, în funcție de criteriile sale, ce diferențe sunt prez-

entate utilizatorului și în ce ordine. Problema identificată este faptul că selecția algoritmică transformă accidental sau intenționat diversitatea într-o experiență personalizată a diversității. Atunci când fiecare persoană primește o versiune diferită și personalizată a lumii, experiența comună devine mai dificil de construit. Nu dispare societatea, însă se modifică infrastructura prin care oamenii ajung să împărtășească semnificații.

Inteligența artificială aduce această problemă la un nivel superior deoarece produce conținut personalizat la o scară fără precedent. Un sistem generativ adaptează formularea unui mesaj, tonul unei conversații, imaginile sau explicațiile la profilul presupus al utilizatorului. Personalizarea produce o formă de „oglindire algoritmică", în care omul primește în permanență o versiune a lumii adaptată propriilor sale caracteristici. În loc ca tehnologia să faciliteze întâlnirea cu ceea ce este diferit, ea o evită.

Problema identității se complică și mai mult atunci când inteligența artificială produce reprezentări ale persoanei. Avatarurile digitale, imaginile generate artificial, vocile sintetice și textele produse automat permit apariția unor forme de reprezentare care nu mai sunt realizate exclusiv de individ. Persoana poate avea o imagine publică generată sau modificată de sisteme automate, iar granița dintre exprimarea autentică și reprezentarea tehnologică devine mai dificil de stabilit.

Această transformare obligă filosofia contemporană să reexamineze conceptul de autenticitate. Din moment ce IA generează în numele unei persoane mesaje, imagini sau opinii, devine necesar să distingem între capacitatea tehnică de reprezentare și consimțământul real al persoanei reprezentate. Totodată, IA afectează și percepția despre Celălalt. Dacă utilizatorul interacționează frecvent cu agenți conversaționali capabili să răspundă instantaneu, să adopte tonuri empatice și să imite forme de dialog uman, apare posibilitatea unei confuzii între relația cu o persoană și relația cu un sistem. Aceasta nu înseamnă că interacțiunea cu inteligența artificială este lipsită de valoare, ci că natura alterității trebuie reconsiderată. Celălalt uman este o ființă autonomă, cu propriile intenții, limite și experiențe, în timp ce sistemul artificial răspunde în baza unei arhitecturi tehnice și a unor modele construite de oameni.

## 5. Concluzii

Cercetarea evidențiază faptul că moderarea algoritmică a conținutului reprezintă una dintre cele mai importante provocări etice, sociale și filosofice ale comunicării în era inteligenței artificiale. Dezvoltarea sistemelor automatizate de identificare, clasificare și filtrare a informației transformă profund modul în care indivizii se exprimă, interacționează și participă la viața publică digitală. Deși aceste tehnologii contribuie la protejarea utilizatorilor împotriva hărțuirii, discursului instigator la ură, dezinformării și altor forme de conținut vătămător, aplicarea lor implică riscul restricționării unor exprimări legitime. Prin urmare, problema fundamentală constă în identificarea unui echilibru între necesitatea protecției colective și respectarea libertății individuale.

Analiza realizată demonstrează că algoritmii reprezintă mecanisme socio-tehnice care participă la organizarea comunicării și la stabilirea limitelor discursului acceptabil. Prin clasificarea și distribuirea conținutului, aceștia influențează vizibilitatea opiniilor, accesul la informație și posibilitățile de participare la spațiul public. Totodată, personalizarea informației favorizează formarea bulelor informaționale, diminuarea contactului cu perspective alternative și fragmentarea experienței sociale comune. În acest context, libertatea de exprimare nu mai presupune exclusiv dreptul de a publica un mesaj, ci și existența unor condiții echitabile pentru ca acesta să ajungă la public.

Din perspectivă antropologică și filosofică, moderarea algoritmică ridică întrebări esențiale privind autonomia individuală, construirea identității și raportarea la alteritate. Clasificarea automată a discursului ignoră particularitățile culturale, intențiile comunicative și semnificațiile contextuale ale mesajelor. În consecință, erorile algoritmice conduc atât la menținerea conținutului nociv, cât și la eliminarea exprimărilor legitime. Aceste limite evidențiază necesitatea recunoașterii complexității limbajului uman și a diversității culturale.

În acest sens, cercetarea susține dezvoltarea unui model de moderare responsabilă, bazat pe transparență, responsabilitate instituțională, supraveghere umană și posibilitatea contestării deciziilor automatizate. Inteligența artificială trebuie utilizată ca instrument de sprijin, fără a înlocui integral judecata umană în situațiile ambigue. În

cele din urmă, legitimitatea moderării depinde de respectarea demnității, autonomiei și pluralismului. Protejarea utilizatorilor și apărarea libertății de exprimare nu trebuie considerate obiective incompatibile, ci principii care se susține reciproc. Numai printr-o guvernanță echilibrată și responsabilă a tehnologiilor digitale se construiește un spațiu public în care siguranța, diversitatea opiniilor și libertatea de exprimare să coexiste.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Referințe bibliografice:**

Thach, H., Mayworm, S., Delmonaco, D., & Haimson, O. (2024), (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. New Media & Society, 26(7).
https://doi.org/10.1177/14614448221109804
Gorwa et al. (2020), Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1).
https://doi.org/10.1177/205395171989
Wright (2022), Automated Platform Governance Through Visibility and Scale: On the Transformational Power of AutoModerator. Social Media and Society.
https://doi.org/10.1177/20563051221077020
Zeng & Kaye (2022), From content moderation to visibility moderation: A case study of platform governance on TikTok. Policy & Internet, 14(1), 79–95.
https://doi.org/10.1002/poi3.287
Christin, A. (2020), Algorithmic ethnography, during and after COVID-19. Sociology Compass, 14(10) https://doi.org/10.1177/20570473209598
Seaver, N. (2017), Algorithms as Culture: Some Tactics for the Ethnography of Algorithmic Systems. Big Data & Society, 4(2), 1–12.
https://doi.org/10.1177/20539517177381
Jereza, R. (2021), Corporeal moderation: Digital labour as affective good. Social Anthropology, 29(4), 928–943. https://doi.org/10.1111/1469-8676.13106

Meppelink, C. S., et al. (2024). It matters how you google it? Using agent-based testing to assess the impact of user choices in search queries and algorithmic personalization on political Google Search results. Journal of Computer-Mediated Communication, 29(6). https://doi.org/10.1093/jcmc/zmae020

**Publisher's Note:** Ius et Societas stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).