ANALIZEAZĂ – DECIDE - ACȚIONEAZĂ

Inteligența Artificială
& Manipularea Percepției Umane

Programul educațional
"Analizează – Decide - Acționează"

Mircea Constantin ȘCHEAU

Alexandru Ciprian ANGHELUȘ

EDUCAȚIA CIBERNETICĂ
FUNDAMENTUL PROTECȚIEI DIGITALE

Într-o lume în care tehnologia evoluează mai rapid decât capacitatea utilizatorilor de a-i înțelege riscurile, educația cibernetică devine unul dintre pilonii siguranței personale și organizaționale. Indiferent dacă discutăm despre utilizatori individuali, angajați, elevi sau lideri decizionali, nu putem să nu luăm în calcul că toți suntem expuși zilnic la amenințări digitale din ce în ce mai sofisticate.

Cunoașterea este armă și armură în același timp.
Prevenția începe cu înțelegerea.

Prin formare continuă, vigilență și aplicare practică a cunoștințelor, putem reduce efectul atacurilor și putem proteja valorile comune: încrederea, intimitatea și integritatea.

Programul educațional ADA - "Analizează – Decide - Acționează"

Gândire critică, responsabilitate digitală și protecție activă în era riscurilor cibernetice

Într-o eră digitală aflată într-o transformare accelerată, unde tehnologia aduce atât oportunități, cât și amenințări, securitatea utilizatorilor devine o prioritate. Atacurile cibernetice moderne exploatează nu doar vulnerabilitățile tehnice, ci mai ales pe cele umane: lipsa de informare, absența vigilenței, supraîncărcarea informațională sau încrederea nefondată în aparențe.
„Analizează – Decide – Acționează” este un program educațional dedicat creșterii rezilienței digitale personale și colective, prin conștientizare, formare practică și dezvoltarea unei gândiri critice adaptate noilor tipuri de agresiuni informaționale.

Seria de materiale din acest program se adresează:

- publicului larg (adulți, seniori, părinți, tineri),
- copiilor și adolescenților (în context educațional),
- funcționarilor publici și profesioniștilor,
- cadrelor de conducere și celor implicați în luarea deciziilor.

Scopul programului este de a transforma informația în instrument de apărare, iar utilizatorul – din țintă pasivă, în actor conștient și activ în fața riscurilor digitale.

Prin aceste materiale ne propunem:

- creșterea gradului de conștientizare și precauție individuală și instituțională;
- diminuarea impactului atacurilor digitale prin educație și reacție rapidă;
- încurajarea culturii de raportare, colaborare și solidaritatea între utilizatori și specialiști;
- formarea unui comportament digital responsabil, în linie cu politicile și reglementările în vigoare.

Proiectul reprezintă un efort susținut de alfabetizare cibernetică, construit pe principii de accesibilitate, aplicabilitate și actualizare constantă, pentru a răspunde dinamicii riscurilor reale din spațiul virtual.

PARTENERI

DIRECTORATUL NAȚIONAL
DE SECURITATE CIBERNETICĂ

Rezumat

Lucrarea explorează moduri în care Inteligența artificială (IA/ AI) este exploatată pentru a influența percepția și pentru a distorsiona realitatea. Fenomenul este analizat prin prisma mecanismelor algoritmice de personalizare, generare de conținut fals credibil (ex: deepfake, voice cloning, text AI-generated etc.), stimulare conversațională și modelare emoțională predictivă.

Scenariile prezentate, în care sunt utilizate tehnici de dezinformare și manipulare ideologică în construcții artificiale sociale, fraude emoționale și atacuri conversaționale automatizate, au rol de exemplificare. Tehnologiile specifice (ex: LLMs, GANs, Emotion AI, feed-uri personalizate, microtargeting) și riscuri sistemice asociate la care se face trimitere, sunt valabile la momentul elaborării prezentului material.

Pe lângă componenta descriptivă, lucrarea oferă un cadru necesar dezvoltării gândirii critice, fiind prezentate metode concrete de detecție, prevenție și reacție, destinate publicului larg, instituțiilor și formatorilor educaționali.

Astfel, documentul se constituie astfel într-un instrument de conștientizare în fața unor forme de inginerie socială automatizată, personalizată și extrem de greu perceptibilă.

Cuvinte-cheie

Inteligență Artificială, manipulare, dezinformare digitală, algoritmi, deepfake, spear phishing, educație, media, securitate.

**Documentul conține termeni tehnici și denumiri standard în mod intenționat, pentru ca toți cititorii să asimileze aceste informații.*

Mesaj către cititori

Inteligența artificială nu este bună sau rea.

Este un instrument. Unul deosebit de puternic. Capabil să învețe din informațiile pe care i le oferim, să reacționeze la stimulii și să producă rezultate în funcție de scopul celor care îl controlează. În mâinile potrivite, AI-ul poate salva vieți, îmbunătăți educația, combate frauda, preveni atacuri cibernetice și susține dezvoltarea societății. În mâini criminale – sau doar iresponsabile – aceeași tehnologie poate fi injectată pentru manipulare, control, înșelătorie, programare ideologică și destabilizare socială.

Tocmai de aceea putem considera că educația este una dintre cele mai potrivite forme de protecție. Dacă înțelegem cum funcționează, putem recunoaște mai ușor când și cum este folosită împotriva noastră. Acest ghid nu are scopul a speria ci doar de a ne pregăti, iar pregătirea începe cu un adevăr simplu:

Inteligența artificială este o unealtă.
Puterea și / sau pericolul sunt în mâinile celor care o folosesc.

Autorii

Manipularea Percepției Umane a trecut prin diferite etape de-a lungul istoriei și în acest sens, Inteligența Artificială se dovedește un instrument eficace și potent în mâinile grupărilor infracționale, tot mai active. O corectă analiză și o bună sursă de informare pot fi ingredientele de succes în efortul legitim de a contracara actorii malițioși, de a limita vulnerabilitățile și de a diminua impactul negativ și pagubele.

Lucrarea este echilibrată, bine structurată și oferă o perspectivă realistă și clară asupra fenomenului. Chiar în contextul unei evoluții rapide a tehnologiilor informaționale, mecanismele de alterare emoțională rămân, totuși, aceleași. Filtrarea cognitivă a conținutului fals poate fi un răspuns în fața avalanșei de algoritmi și boți conversaționali. Antrenarea gândirii critice digitale este absolut necesară pentru a evita profilarea comportamentală predictivă și modelarea malițioasă a deciziilor utilizatorului.

Analizele sunt destinate publicului larg, profesioniștilor și profanilor deopotrivă. Ghidul pune accent pe latura practică a problematicilor și face trimitere la surse oferite de organisme și instituții relevante, cu atribuții în domeniu. Efortul comun a condus la elaborarea prezentului studiu, unul dintre deziderate fiind creșterea rapidă a nivelului de conștientizare și prin aceasta, consolidarea culturii de securitate cibernetică și a rezilienței.

Directoratul Național de Securitate Cibernetică (DNSC)

Criminalitatea informatică se află într-o continuă evoluție, pe fondul dezvoltării accelerate a tehnologiilor digitale și al extinderii utilizării internetului în viața de zi cu zi. Aceste transformări aduc beneficii semnificative, dar și riscuri sporite pentru utilizatori.

Inteligența Artificială (IA) are un impact tot mai pronunțat asupra societății, influențând modul în care interacționăm în mediul digital. Înțelegerea mecanismelor de bază ale IA și a modului în care aceasta poate fi folosită în scopuri frauduloase contribuie esențial la prevenirea criminalității informatice.

Prezentul ghid, elaborat în sprijinul activităților de prevenire, se adresează atât publicului larg, cât și profesioniștilor, oferind repere utile pentru o utilizare responsabilă și sigură a tehnologiilor digitale.

Institutul de Cercetare și Prevenire a Criminalității

CUPRINS

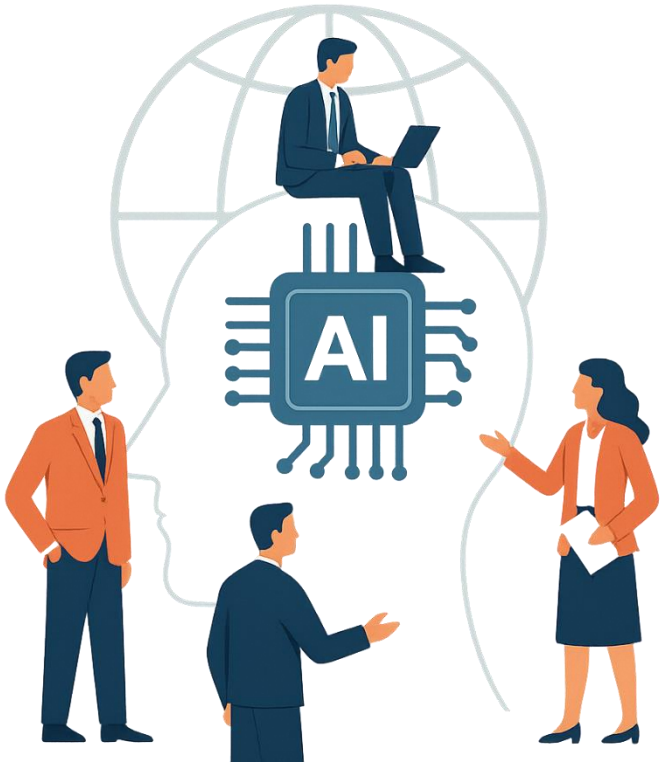
1. NOȚIUNI GENERALE

1.1 Ce este Inteligența Artificială

Inteligența artificială (IA) reprezintă un ansamblu de tehnologii informatice capabile să simuleze procese de gândire umană – precum învățarea, raționamentul, percepția și luarea deciziilor. Cele mai cunoscute tipuri includ: învățarea automată (machine learning), rețelele neuronale profunde (deep learning), procesarea limbajului natural (NLP), recunoașterea vizuală, precum și modelele generative (ex: ChatGPT, DALL·E, Gemini, Claude etc.).

Spre deosebire de algoritmii clasici, IA modernă nu urmează un set fix de reguli, ci „învață" din seturi de date și se adaptează comportamentului uman. Modelele avansate, precum cele generative, pot crea texte, imagini, voci sau chiar videoclipuri aproape imposibil de deosebit de cele reale.

Aceste capacități oferă beneficii extraordinare – de la automatizare la educație și cercetare – dar implică și riscuri semnificative, în special în sfera manipulării informației, încrederii și emoțiilor umane.



1.2 Percepția umană și Inteligența Artificială

În literatura de specialitate din limba engleză Inteligența Artificială se traduce prin Artificial Intelligence și de aceea se va utiliza în text acronimul AI.

Inteligența artificială (AI) nu mai este doar un instrument al viitorului – este o parte activă a prezentului nostru digital. Fie că vizionăm un videoclip pe o platformă de streaming, citim o știre online sau purtăm o conversație cu un asistent virtual, există o probabilitate destul de ridicată ca în fundal să ruleze un algoritm de inteligență artificială care decide ce vedem, ce auzim și chiar cum ar trebui să interpretăm realitatea interpretăm realitatea.

La baza acestor procese se află capacitatea AI-ului de a analiza comportamente umane, de a învăța din datele colectate și de a genera conținut sau răspunsuri care imită sau stimulează reacțiile umane autentice. Aceste caracteristici au aplicații valoroase în medicină, educație sau automatizare, dar există și un revers periculos: riscul de manipulare informațională și emoțională pe scară largă.

Spre deosebire de metodele clasice de influențare (ex: publicitate, propagandă, persuasiune socială), AI introduce un nou nivel de precizie și discreție în manipulare. Algoritmii moderni pot identifica punctele slabe emoționale, tiparele comportamentale și preferințele psihologice ale utilizatorilor cu o acuratețe uluitoare, generând conținut personalizat care activează emoții puternice și decizii impulsive – fără ca victima să conștientizeze acest proces.

De la recomandări de conținut, care întăresc convingerile proprii și izolează utilizatorii în bule informaționale, până la simularea cu mare acuratețe a unor persoane reale (prin deepfake, voice cloning sau avataruri AI), inteligența artificială devine un vector activ de modelare a percepției și, implicit, al realității subiective a fiecărui individ.

Această nouă realitate ridică întrebări pertinente:

- Cum recunoaștem un conținut generat sau manipulat de AI?
- Care este limita între recomandare utilă și manipulare intenționată?
- Ce înseamnă adevăr într-o epocă în care orice voce, imagine sau activitate poate fi replicată cu precizie artificială?

Pentru a răspunde acestor provocări, este necesară o educație cibernetică extinsă, care să depășească noțiunile de bază și să abordeze dimensiunea cognitivă, psihologică și socială a interacțiunii cu tehnologiile inteligente.

Acest ghid educațional își propune să:

- Explice modalități în care AI poate influența percepția umană, prin mecanisme vizibile sau subtile;
- Analiza de riscuri și tehnologii, oferind o privire cât mai realistă asupra capabilităților curente ale AI-ului în domeniul manipulării cognitive;
- Ofere metode concrete de prevenție, verificare și apărare, atât pentru utilizatorii individuali, cât și pentru organizații sau instituții expuse riscurilor informaționale;
- Contribuie la formarea unei gândiri critice digitale – o componentă importantă pentru navigarea într-un spațiu digital tot mai personalizat, influențat și potențial manipulativ.

1.3 De ce este importantă înțelegerea mecanismelor din spatele AI-ului

Pe măsură ce inteligența artificială devine tot mai integrată în viața cotidiană, înțelegerea modului în care aceste sisteme funcționează nu mai este doar o preocupare a specialiștilor în tehnologie. Este nevoie de acest tip de înțelegere pentru orice utilizator de internet, pentru decidenți, educatori sau părinți. Interacționăm zilnic cu aplicații alimentate de algoritmi inteligenți, dar de cele mai multe ori nu avem conștiența proceselor care se desfășoară „în spate”.

Această lipsă de transparență creează un dezechilibru major: sistemele știu foarte multe despre noi, în timp ce noi știm foarte puțin despre ele. În timp ce AI-ul colectează date, analizează comportamente și ajustează mesajele livrate în funcție de profiluri psihologice, utilizatorul obișnuit rămâne într-o poziție pasivă, cu o percepție de control care este adesea iluzorie.

În acest context, educația cibernetică trebuie să includă și o componentă de alfabetizare tehnologică: înțelegerea mecanismelor de bază care fac posibilă personalizarea conținutului,

recunoașterea emoțiilor, predicția comportamentului sau generarea automată de conținut aparent uman.

Această înțelegere nu presupune competențe avansate de programare sau matematică, ci curiozitate și spirit critic:

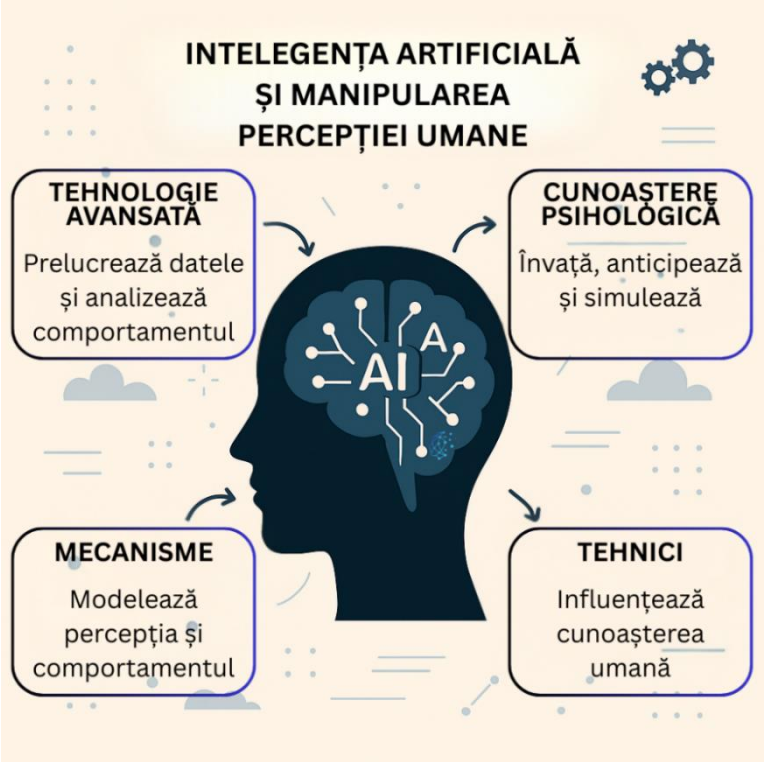
- Ce este un algoritm de recomandare și cum influențează ceea ce văd?
- Cum poate o rețea neuronală să înțeleagă emoțiile mele?
- Ce se întâmplă cu datele pe care le ofer, conștient sau inconștient?
- De ce anumite conținuturi par „făcute exact pentru mine”?

Răspunsul la aceste întrebări permite o schimbare de paradigmă: din simplu consumator digital pasiv, utilizatorul devine un actor conștient, capabil să-și gestioneze expunerea, să pună întrebări și să reacționeze informat. Fără acest filtru, riscul este de a trăi într-o realitate percepută, modelată de mașini, fără o minimă capacitate de reflecție sau de verificare.

Capitolele următoare vor detalia aceste mecanisme tehnice și vor arăta cum algoritmii pot capta, influența sau distorsiona percepția, de cele mai multe ori fără a lăsa urme evidente.

2 ELEMENTE TEHNICE ȘI MECANISME

Manipularea percepției umane cu ajutorul inteligenței artificiale se bazează pe o sinergie între tehnologiile avansate și cunoaștere psihologică. AI-ul nu este doar un instrument de procesare a datelor – este chiar un sistem care învață, anticipează, influențează și simulează comportamente umane cu precizie în creștere accelerată.



În această secțiune vom analiza tehnologiile care stau la baza influenței perceptive și mecanismele operaționale prin care acestea sunt exploatate pentru a modela percepția și comportamentul uman.

Pentru a înțelege cum aceste sisteme ajung să ne influențeze percepțiile, este important să știm, chiar și la nivel introductiv cum funcționează modelele de inteligență artificială. Nu este vorba despre detalii matematice sau algoritmice complexe, ci despre o înțelegere funcțională: ce fac aceste modele, cum învață și cum pot simula comportamente inteligente.

La baza celor mai avansate aplicații AI se află așa-numitele rețele neuronale artificiale – structuri matematice inspirate din modul de funcționare al creierului uman. Acestea sunt alcătuite din multiple straturi de „neuroni artificiali" care procesează informația treptat, extrăgând semnificații din datele primite (cum ar fi text, imagine, sunet). Fiecare strat filtrează, interpretează și transmite mai departe informația, până când sistemul produce un rezultat.

Modelele AI moderne sunt antrenate pe cantități uriașe de date. În timpul acestui proces, ele „învață" să recunoască tipare, relații, emoții sau intenții. Cu cât datele sunt mai variate și cu cât procesul de învățare este mai bine calibrat, cu atât rezultatul devine mai precis și mai credibil.

Un exemplu simplu este modelul care îți recomandă conținut pe o platformă de video sau social media. Acesta observă ce fel de materiale urmărești, cât timp petreci pe ele, cum reacționezi (like, comentariu, distribuire) și, în timp, îți livrează conținut tot mai adaptat intereselor tale – chiar și fără să-i fi spus explicit ce preferi.

Modelele de inteligență artificială pot fi clasificate, într-un mod simplificat, în funcție de scopul și complexitatea lor:

- Machine Learning (ML) - modele care învață din date pentru a face predicții sau clasificări. Exemple: recunoașterea unui spam, recomandarea unui produs.
- Deep Learning (DL) - o formă avansată de ML, care folosește rețele neuronale profunde și poate identifica tipare foarte complexe, cum ar fi tonul emoțional dintr-o voce sau intenția dintr-un text.
- Modele generative - capabile să creeze conținut nou – texte, imagini, videoclipuri sau voci – care pot părea autentice. Exemple: ChatGPT, DALL·E, voice cloning.
- AI conversational - proiectat pentru a purta dialoguri fluente și convingătoare, adaptate emoțional la utilizator.

Pe măsură ce aceste modele devin mai avansate, ele nu doar reacționează la utilizator, ci încep să îl modeleze activ: pot influența deciziile, pot simula empatie, pot anticipa reacții și pot livra conținut care vizează direct slăbiciunile sau preferințele personale.

2.1 Tehnologii implicate

A. Rețele neuronale artificiale (deep learning)

„Creierul digital" care învață, creează și simulează comportamente umane.

Influențarea percepției prin inteligență artificială se bazează pe un ecosistem de tehnologii interconectate care simulează, anticipează și modelează comportamentele umane. Aceste sisteme nu doar reacționează la inputuri, ci intervin activ în modelarea realității percepute de utilizatori – uneori în mod subtil, alteori în mod direct. Mai jos sunt prezentate principalele clase de tehnologii implicate în acest proces, cu accent pe mecanismele lor de funcționare și riscurile aferente.

Rețelele neuronale artificiale reprezintă coloana vertebrală a inteligenței artificiale moderne. Acestea sunt sisteme de învățare automată inspirate din modul în care funcționează creierul uman – mai precis, din structura și comportamentul neuronilor biologici.

Un sistem de deep learning este format din straturi multiple de noduri (neuroni artificiali), care primesc informație, o procesează, și apoi o transmit mai departe în rețea. Prin ajustarea conexiunilor între acești „neuroni", sistemul învață din date și devine tot mai precis în recunoașterea tiparelor sau generarea de conținut. Cum funcționează, în esență?

- Sistemul este „hrănit" cu date (ex: mii de imagini, fragmente de text, înregistrări audio);
- Fiecare strat al rețelei extrage caracteristici tot mai abstracte (ex: de la pixeli → forme → expresii faciale);
- Pe măsură ce învață, sistemul își „optimizează" conexiunile pentru a prezice, clasifica sau genera date noi;
- După o perioadă de antrenament, rețeaua poate răspunde la stimuli complet noi – cu o precizie apropiată de comportamentul uman.

Această arhitectură complexă face ca rețelele neuronale să fie utilizate într-o gamă largă de aplicații – de la recunoaștere facială la traduceri automate, de la sisteme medicale la platforme de divertisment. Dar una dintre cele mai influente și mai controversate direcții este modelarea percepției umane.

A1. Aplicabilități în manipularea percepției.

Puterea rețelelor neuronale nu constă doar în analiză, ci și în capacitatea lor de a interveni asupra emoțiilor și convingerilor umane. Printre cele mai relevante utilizări se numără:

Analiza expresiilor faciale

Rețelele pot detecta micro-expresii, emoții subtile și stări afective prin analiză video. Sunt folosite în:

- Reclame personalizate,
- Analiză de interviuri,
- Manipulare emoțională automată (ex: aplicații de dating, educație adaptivă).
- Generarea de conținut – text, video, audio

Cu ajutorul rețelelor neuronale, AI poate:

- Genera texte convingătoare (ex: articole, postări, știri false),
- Crea videoclipuri deepfake,
- Sintetiza voci umane realiste pentru escrocherii sau persuasiune.

Exemplu: un mesaj vocal „primit" de la o rudă sau superior, generat complet de AI.

Recunoașterea emoțiilor

Prin analizarea comportamentului digital (ex: voce, mimică, ritm de tastare), AI-ul identifică starea emoțională a utilizatorului și adaptează reacțiile sale pentru a:

- Întări loialitatea,
- Declanșa reacții impulsive,
- Manipula decizii în momente de vulnerabilitate

Sinteza naturală a vocii

Folosind rețele specializate (ex: Tacotron, WaveNet, HiFiGAN), AI poate crea voci complet sintetice care:

- Imită o persoană reală (ex: voice cloning),
- Exprimă emoții autentice,
- Comunică mesaje persuasive cu intonație, pauze și ritm natural.

Riscuri și implicații

- Capacitate ridicată de generare realistă extremă – poate crea conținut fals imposibil de identificat vizual sau auditiv de către utilizatori obișnuiți;
- Automatizarea manipulării psihologice – AI învață cum să reacționeze la emoțiile umane mai eficient decât un om neantrenat;
- Escaladarea atacurilor din zona ingineriei sociale – atacatorii pot folosi rețele neuronale pentru a declanșa atacuri personalizate pe scară largă (ex: fraude, manipulare politică, șantaj).

Conștientizare

Rețelele neuronale sunt fundamentale pentru AI, dar puterea lor de influențare este direct proporțională cu ignoranța publicului. Cu cât înțelegem mai bine cum funcționează și ce efecte se pot înregistra, cu atât mai mult putem:

- Adresa întrebări critice când interacționăm cu conținut aparent convingător,
- Refuza impulsurile generate algoritmic,
- Susține reglementarea responsabilă a acestor tehnologii.

A2. Sisteme de recomandare și predicție comportamentală

"Mașinile care știu ce vei face – uneori, mai bine decât tine."

Sistemele de recomandare și predicție comportamentală sunt printre cele mai utilizate și în același timp, cele mai puțin înțelese componente ale inteligenței artificiale moderne de către publicul larg. Ele sunt prezente și acționează în fundalul aproape tuturor interacțiunilor noastre digitale – de la videoclipuri pe YouTube, până la produse ce apar în feed-ul de cumpărături sau postări de pe rețelele sociale.

Aceste sisteme folosesc AI și machine learning pentru a analiza comportamentul online și a construi un model predictiv despre utilizator. Scopul declarat este de a îmbunătăți experiența digitală. Însă în practică, aceste sisteme pot fi deturnate pentru a influența, manipula și controla deciziile, emoțiile și convingerile utilizatorilor, fără ca aceștia să conștientizeze și să-și exprime acordul explicit.

Cum funcționează?

Sistemele de recomandare colectează și analizează informații despre tine, precum:

- Istoricul de căutare și navigare,
- Durata de vizualizare a unui conținut,
- Clicuri, like-uri, comentarii, partajări,
- Ritmul de derulare, pauzele și revenirea pe conținut,
- Locație, ora din zi, dispozitiv folosit, comportamente repetitive.

Aceste date sunt procesate pentru a crea un profil comportamental și, ulterior, un set de previziuni personalizate (ex: ce te interesează, ce te atrage, ce te neliniștește, ce tip de reacție este cel mai probabil să ai etc.).

Ce pot face aceste sisteme?

- Pot alege ce vezi și în ce ordine vezi – feed-urile nu sunt cronologice, ci modelate să mențină atenția ta cât mai mult;
- Anticipează ce vei face – dacă ești pe cale să cumperi, să distribui ceva sau să te enervezi, sistemul „știe" și acționează în consecință;
- Te influențează emoțional – prin livrarea de conținut care stârnește reacții afective intense (ex: teamă, furie, dorință, revoltă);

- Îți modelează obiceiurile – prin repetiție și expunere strategică, ajungi să adopți anumite rutine digitale fără să-ți dai seama.

Exemple concrete:

- Un utilizator vizionează clipuri legate de sănătate → sistemul începe să-i recomande produse „naturiste" scumpe sau conținut conspiraționist;
- O persoană comentează un articol de natură politică → primește postări din ce în ce mai partizane, care întăresc doar propria viziune (sau o deturnează);
- Cineva caută „cum să gestionez stresul" → este inundat cu reclame pentru cursuri costisitoare, aplicații de relaxare și soluții rapide, uneori inutile;
- O adolescentă urmărește conținut legat de greutatea corporală → sistemul începe să recomande materiale care promovează idealuri toxice de frumusețe sau comportamente alimentare riscante.

Riscuri majore:

- Manipulare invizibilă a convingerilor – utilizatorul ajunge să creadă că ideile și deciziile sale îi aparțin în totalitate, când de fapt sunt induse prin expunere algoritmică repetitivă;
- Polarizare socială și ideologică – când oamenii sunt "injectați" doar cu opinii similare, diferențele de viziune devin extreme, de neînțeles și de neacceptat;
- Comportament modelat pe obiective comerciale – nu vezi ceea ce ai nevoie, ci ceea ce are cea mai mare valoare economică sau ideologică pentru platformă sau promotor;
- Dependență digitală – sistemele optimizează pentru atenție, nu pentru echilibru. Astfel, oferă conținut stimulant și creator de dependență, nu util și sănătos.

Cum ne putem proteja?

- Folosim platformele în mod conștient, nu în regim pasiv (ex: derulare automată, consum excesiv);
- Setăm limite de timp și opțiuni de personalizare unde este posibil;
- Căutăm activ surse și conținut alternativ, nu doar ce ni se oferă în feed;
- Folosim periodic sesiuni de navigare curată (ex: incognito, fără login, fără istoric activat);
- Ne întrebăm frecvent: „Cine a ales să văd asta? Eu sau un algoritm?"

B. Large Language Models (LLMs) (ex: ChatGPT, Gemini, Claude, Mistral etc)

„Inteligența Artificială care înțelege și generează limbajul uman – cu precizie, viteză și influență fără precedent."

Large Language Models (LLMs) reprezintă o clasă de algoritmi AI antrenați pe cantități masive de text – zeci, sute de miliarde sau trilioane de cuvinte, extrase din cărți, articole, conversații, site-uri web, coduri de programare și multe altele. Aceste modele au capacitatea de a înțelege, interpreta, simula și genera limbaj uman coerent, adaptat contextului, publicului și intenției dorite.

LLM-urile precum ChatGPT (OpenAI), Gemini (Google), Claude (Anthropic), Mistral, LLaMA (Meta) sau Command-R (Cohere) sunt deja integrate în numeroase aplicații comerciale, educaționale, organizaționale și sociale.

Cum funcționează?

- Modelul este antrenat pe seturi incredibil de mari de date (ex: corpusuri lingvistice globale);

- Învățarea are loc prin predicția cuvântului următor într-un context dat – dar cu milioane de exemple;
- După antrenare, AI-ul poate răspunde la întrebări, redacta texte, rezuma informații, construi argumente, conversa pe diverse teme și chiar simula stiluri și emoții.

Capacități relevante în manipularea percepției:

- Generare automată de text persuasiv – articole, opinii, știri false, argumentații convingătoare;
- Răspunsuri aparent neutre, dar ideologic influențate, în funcție de sursele de antrenament și parametrii;
- Simulare de voci online (ex: text-based impersonation) – un AI poate răspunde ca și cum ar fi o persoană reală în baza unui text;
- Asistență conversațională manipulativă – ghidarea subtilă a utilizatorului spre anumite concluzii, produse, idei.

Exemple concrete:

- Un actor malițios folosește un LLM pentru a genera sute de articole „expert" referitoare la o temă controversată – toate având aceeași direcție ideologică;
- Un chatbot de asistență aparent empatic sugerează unui utilizator vulnerabil „soluții" care conduc spre achiziții riscante sau convingeri toxice;
- Un model AI este configurat să răspundă „calm și profesionist" în contexte de dezinformare, oferind mesaje persuasive falsificate.

Riscuri și implicații:

- Scalabilitate infinită a manipulării – un LLM poate produce în câteva minute mii de texte adaptate emoțional, ideologic sau contextual pentru a influența publicul țintă;
- Deghizare în autoritate – dacă un AI este „mascat" ca expert, profesor, consilier sau lider, mesajele sale pot influența decizii majore;
- Imposibilitatea de a distinge conținutul uman de cel generat – textele par naturale, coerente, credibile – chiar dacă sunt complet fabricate;
- Automatizarea ingineriei sociale – un LLM poate învăța și aplica tactici clasice de persuasiune, manipulare și dezinformare, la scară largă și fără pauză.

Cum ne protejăm?

- Folosim LLM-urile ca instrumente, nu ca surse "absolute" de adevăr;
- Verificăm sursa inițială / primară a informației, mai ales când pare „prea bine scrisă" sau „perfect argumentată";
- Dezvoltăm sănătos reflexul de a întreba: este acest conținut generat de om sau de AI?;
- Recunoaștem modele persuasive: repetiție subtilă, apel la emoții, ton hiper rațional, evitarea surselor verificate.

C. Machine Learning Afectiv (Emotion AI) – detectarea și manipularea emoțiilor

„Când AI-ul nu doar te ascultă sau te privește, ci te simte și reacționează în consecință."

Emotion AI, cunoscut și sub denumirea de machine learning afectiv, se individualizează ca o ramură a inteligenței artificiale ce are, printre altele, rolul de a detecta, interpreta și reacționa la stările emoționale ale oamenilor. Spre deosebire de AI-ul tradițional care procesează date explicite (ex: cuvinte, comenzi, cifre), Emotion AI se concentrează pe date implicite, subtile și contextuale, precum expresiile faciale, tonul vocii, ritmul respirației sau comportamentul digital.

Această tehnologie face ca AI-ul să nu mai fie un simplu „executant de comenzi”, ci un interlocutor sensibil la emoțiile umane, capabil să reacționeze empatic sau să exploateze emoții în scenarii periculoase pentru a manipula.

Cum funcționează?

- Colectare de semnale afective – prin cameră video, microfon, tastatură, mouse, senzori biometrici sau analiza comportamentului digital;
- Analiză multimodală – combinarea diferitelor tipuri de date (ex: voce + expresie facială + activitate digitală) pentru o înțelegere holistică a stării emoționale;
- Modelare și interpretare – AI-ul clasifică starea utilizatorului ca: stresat, trist, euforic, furios, anxios etc.;
- Reacție adaptivă – AI-ul ajustează conținutul, tonul sau ritmul interacțiunii în funcție de emoția detectată.

Unde este folosit?

- Aplicații de relații clienți (ex: customer service): chatbot-ul „se adaptează” în funcție de tonul tău;
- Educație digitală adaptivă: detectează frustrarea sau plictiseala și schimbă strategia de învățare;
- Publicitate țintită emoțional: reclame afișate când AI-ul „simte” că ești vulnerabil sau receptiv;
- Securitate și supraveghere: recunoaștere facială emoțională în aeroporturi, școli, stadioane;
- Asistență psihologică AI: conversații emoțional-reactive cu utilizatori în dificultate.

Exemple de manipulare:

- AI-ul detectează anxietate și livrează reclame cu mesaj alarmist: „Ești pregătit pentru ce urmează?”;
- AI-ul „simte” tristețea și direcționează utilizatorul spre conținut de consolare – inclusiv mesaje care manipulează sau exploatează emoțional;
- În contexte comerciale, AI-ul recunoaște frustrarea și propune „soluții” cu costuri mari sau condiții dezavantajoase;
- În propagandă, algoritmii oferă conținut ideologic intens, exact în momentele emoționale de vulnerabilitate cognitivă.

De ce este periculos?

- Exploatează momente de instabilitate emoțională – când ești supărat, anxios sau entuziasmat, ești mai ușor de manipulat;
- Este invizibil și imposibil de verificat pentru utilizator – nu știi când ești „evaluat emoțional”, nici ce face sistemul cu acea informație;
- Permite controlul psihologic automatizat – AI-ul ajustează vocea, mesajul, culorile, muzica sau dinamica unei interacțiuni pentru a stimula reacții precise (ex: acceptare, teamă, impuls, achiziție, evitare etc.).
- Încalcă intimitatea mentală – este una dintre cele mai directe forme de intruziune în spațiul psihologic personal, fără acord conștient.

Cum ne putem proteja?

- Limităm accesul aplicațiilor la cameră video, microfon, senzori și date biometrice, dacă nu este necesar;
- Folosim software-uri sau extensii care reduc monitorizarea comportamentală;

- Recunoaștem momentele noastre de slăbiciune emoțională și evităm să luăm atunci decizii importante;
- Ne întrebăm: „Reacționez pentru că simt cu adevărat asta – sau pentru că un sistem m-a condus acolo?”.

D. AI vizual – imagini, video, deepfake, avataruri sintetice

„Când ceea ce vezi cu ochii tăi poate fi complet fals – dar imposibil de detectat.”

AI-ul vizual este zona în care inteligența artificială procesează, înțelege și generează conținut vizual: imagini statice, videoclipuri, animații și chiar entități virtuale sintetice care interacționează cu utilizatorii. Este una dintre cele mai spectaculoase, dar și cele mai periculoase direcții ale tehnologiei AI, cu impact direct asupra sentimentului de încredere vizuală – acel instinct fundamental uman de a crede ce vedem.

De la simple îmbunătățiri de imagine, la reconstrucții faciale complete, de la generarea unor p care nu există, până la manipularea expresiilor faciale în timp real, AI-ul vizual redefinește conceptul de autenticitate vizuală.

Cum funcționează?

- Modele AI antrenate pe seturi de date vizuale învață să recunoască, genereze și modifice imagini și videoclipuri.
- Se folosesc algoritmi precum:
    - GANs (Generative Adversarial Networks) – pentru generare de imagini realiste;
    - Autoencoders – pentru reconstrucția trăsăturilor faciale;
    - Deepfake frameworks – pentru înlocuirea feței sau sincronizarea buzelor;
    - Text-to-image models – pentru crearea de imagini sintetice din descrieri text (ex: Midjourney, DALL·E, Stable Diffusion);
    - Motion capture AI – pentru animarea avatarurilor în timp real.

Capacități și aplicații:

- Crearea de indivizi sau grupuri care nu există, dar care par absolut reale (ex: fotografii, profiluri sociale, influenceri virtuali);
- Deepfake video/audio – înlocuirea feței și vocii unei persoane reale dintr-un material video, pentru a simula o declarație, un gest sau o acțiune;
- Avataruri interactive AI – personaje animate, cu fețe sintetice, care interacționează conversațional cu utilizatorul;
- Modificarea expresiilor și a emoțiilor din materiale vizuale existente, fără a altera peisajul natural al scenei.

Exemple concrete de manipulare:

- Un videoclip în care un politician pare să recunoască și chiar să-și asume fapte grave – dar declarația nu a fost niciodată rostită;
- O campanie de dezinformare în care „martori oculari” comentează un eveniment – deși persoanele nu există, fiind create integral de AI;
- Un mentor (influencer) „virtual” care promovează produse, ideologii sau cauze, dar în realitate este doar un construct digital controlat de o agenție;
- O aplicație de tip filtru fotografic (beauty filter AI) care schimbă subtil trăsăturile utilizatorilor în scopuri comerciale și de control al percepției de sine.

De ce este periculos?

- Fractura dintre realitate și aparență – ceea ce este vizibil nu mai este un indicator de încredere și ochiul uman nu mai poate distinge falsul de autentic fără instrumente de verificare specializate;
- Manipularea emoțională la nivel profund – imaginile și video-urile sunt procesate emoțional mai rapid și mai intens decât textul și un fals vizual bine articulat declanșează reacții automate;
- Efect de contagiune socială – materialele deepfake sau synthmedia pot deveni virale foarte rapid, producând reacții publice masive înainte de a fi verificate;
- Abuz, șantaj, dezinformare, compromitere personală sau instituțională – prin falsificarea conținutului vizual, reputații și relații pot fi distruse instantaneu.

Cum ne putem proteja?

- Verificăm autenticitatea vizuală cu unelte specializate:
  - Deepware Scanner,
  - Sensity AI,
  - Microsoft Video Authenticator,
  - InVID verification plugin (pentru jurnaliști).
- Căutăm sursa originală a materialului vizual: unde a fost publicat prima dată, de cine, în ce context?
- Rămânem critici în fața conținutului „șocant" sau „senzațional": dacă pare prea real sau prea extrem și reclamă a fi verificat.
- Raportăm imediat falsurile periculoase pe platformele unde apar și alertăm comunitatea și autoritățile.

2.2 Mecanisme de manipulare perceptivă

Dincolo de tehnologie, manipularea eficientă a percepției implică înțelegerea profundă a psihologiei umane. Mecanismele utilizate de AI sunt concepute pentru a exploata reacții emoționale, biasuri cognitive(capcane mentale) și tipare comportamentale recurente.

A. Algoritmi de știri (feed) personalizat

„Ce vezi nu e întâmplător – ci programat să te captiveze și să te influențeze."

Feed-urile personalizate au devenit axul central al experienței digitale moderne. Când navighezi pe o rețea socială, citești știri online, cauți informații sau vizionezi videoclipuri, nu vezi tot ceea ce există, ci doar ceea ce algoritmul decide să îți arate. Această decizie nu este neutră și nu are ca scop diversitatea sau echilibrul informațional, ci maximizarea implicării tale – adică a atenției, emoțiilor și reacțiilor tale.

Cum funcționează?

Algoritmii de feed personalizat utilizează inteligență artificială pentru a analiza:

- ce tip de conținut consumi cel mai des,
- cât timp petreci parcurgând un anumit articol, postare sau video,
- ce distribui, comentezi sau apreciezi,
- cine sunt persoanele și paginile cu care interacționezi,
- la ce oră, de pe ce dispozitiv accesezi și în ce stare emoțională te afli (în funcție de comportament și semnale subtile).

În baza acestor informații, AI-ul construiește un profil comportamental și emoțional și îți servește un feed structurat pe măsură: conținut care are șanse mari să declanșeze o reacție imediată.

Ce fel de conținut este afișat?

- Informații care confirmă convingerile tale
  - Dacă ai dat like(apreciat) sau ai comentat un articol anti-vaccin, îți vor apărea și altele similare, nu argumente pro-vaccin;
  - Dacă ești interesat de o anumită ideologie, AI-ul îți oferă postări care o susțin, nu critici obiective.
- Postări care activează emoții intense
  - Frică, furie, indignare, admirație – orice emoție care determină o reacție rapidă;
  - Algoritmii favorizează conținutul „viral”(cu impact) emoțional, nu pe cel echilibrat și informativ.
- Perspective identice cu ale tale
  - „Toată lumea” pare să gândească exact ca tine.
  - Dispar opiniile contrare, punctele de vedere divergente, argumentele alternative.

Exemplu concret:

Un utilizator cu simpatii politice solide începe să acceseze postări în care se critică o anumită idee sau categorie socială. În scurt timp:

- Feed-ul său devine dominat de mesaje similare, unele chiar extreme;
- Postările moderate dispar sau sunt foarte rare;
- AI-ul accentuează „ideea dominantă” pentru a-l menține conectat și implicat;
- Utilizatorul are impresia că opinia sa este unanim acceptată și susținută.

Rezultat: radicalizare, polarizare, izolare informațională

- Radicalizare: opiniile se întăresc în lipsa dezbaterii și a diversității informaționale;
- Polarizare: taberele devin tot mai rigide și mai intolerante una față de cealaltă;
- Izolare: utilizatorii trăiesc în „camere de ecou” algoritmice unde se aud doar propriile idei, amplificate.

Efecte asupra percepției:

- Realitatea devine distorsionată – dacă vezi doar dintr-un unghi de abordare, îl percepi ca fiind „adevărul absolut”;
- Dezbaterea publică devine toxică – lipsa expunerii la argumente contrare conduce la refuzul dialogului;
- Societatea devine fragmentată – fiecare grup trăiește într-o realitate paralelă, modelată de algoritmi.

Cum ne putem proteja?

- Diversificăm sursele de informare – urmărim conștient și perspective opuse / adverse:
- Căutăm activ conținut din afara feed-ului algoritmic – accesăm direct surse de știri independente, canale specializate, pagini nepersonalizate:
- Limităm timpul petrecut în platforme care nu oferă control asupra feed-ului;
- Ne întrebăm constant: „Cine a ales să văd asta? Eu – sau o mașină care știe ce mă interesează cel mai mult?”.

B. Microtargeting psihografic – influențarea personalizată a percepției și comportamentului

„Când AI-ul îți cunoaște fricile, speranțele și slăbiciunile – și le folosește împotriva ta.”

Microtargeting-ul psihografic este o tehnică avansată de influențare digitală ce combină analiza comportamentală, psihologică și emoțională cu inteligența artificială, pentru a crea mesaje personalizate și direcționate individual, cu scopul de a influența convingerile, deciziile și comportamentul.

Spre deosebire de publicitatea clasică, ce transmite un mesaj general către un public larg, microtargeting-ul AI creează campanii personalizate pentru fiecare individ sau grup de indivizi – adaptate stilului cognitiv, emoțiilor predominante, vulnerabilităților și contextului social.

Cum funcționează?

- AI-ul colectează și analizează date despre subiect din surse multiple:
  - like-uri, share-uri, comentarii, istoricul de căutări și achiziții;
  - postări scrise, limbaj folosit, emoticoane, orar de activitate;
  - locație, contacte, grupuri, preferințe politice;
  - date demografice și semnale comportamentale.
- Pe baza acestor informații, construiește un profil psihografic detaliat:
  - ce motivează, ce sperie, cum se reacționează la autoritate, ce stil de comunicare se preferă, în ce tip de mesaje se investește încredere.
- Apoi, livrează mesaje personalizate, prin reclame, postări, articole, videoclipuri sau conversații simulate, pentru a:
  - influența achiziția un produs,
  - convinge (re)orientarea pentru a se susține o cauză,
  - exprima votul într-un anumit fel,
  - convinge (re)orientarea pentru respingerea unui grup sau a unei idei.

Exemple concrete:

- O persoană cu tendințe anxioase primește conținut care accentuează riscuri, crize, soluții de urgență;
- Un utilizator cu orientare politică ambiguă este expus la argumente subtile, dar repetate, menite să-l atragă într-o anumită tabără;
- Un adolescent cu stimă de sine scăzută este bombardat cu reclame pentru produse de transformare fizică, aplicații de dating(întâlniri) sau comunități „exclusive”;
- Un angajat care își exprimă nemulțumiri online primește invitații la mișcări de protest sau grupuri ideologice radicale.

De ce este periculos?

- Elimină autonomia decizională – deciziile devin doar reacții la mesaje concepute special pentru a influența;
- Este complet invizibil – nu știm că am fost targetați(vizați) și nici nu realizăm că ceea ce ni se oferă este personalizat cu scop de manipulare;
- Funcționează tăcut și fără opoziție – pentru că mesajul pare „logic” sau „natural”, nu declanșează scepticism sau reacție critică;
- Poate radicaliza – utilizatorii care nu au acces la alte puncte de vedere sunt ușor de atras spre extreme ideologice.

Exemple de impact major:

- Campania Cambridge Analytica – unde milioane de alegători au fost influențați prin microtargeting psihografic în procesele de alegeri;
- Campanii anti-vaccinare ce se foloseau de teamă, incertitudine și neîncredere în autorități, țintind segmente vulnerabile emoțional;
- Platforme comerciale care vând produse de slăbit sau „soluții miraculoase" doar către persoane cu tipare detectate de insecuritate fizică sau depresie latentă.

Cum ne putem proteja?

- Limităm cantitatea de date personale partajate pe rețele sociale și platforme online;
- Folosim extensii și setări care blochează tracking-ul comportamental (ex: Privacy Badger, uBlock Origin, Ghostery);
- Folosim conturi „curate" pentru căutări importante (ex: în browsere nepersonalizate/incognito);
- Analizăm critic mesajele care „par să fie exact pentru noi" – și care sunt de fapt cele mai suspecte.

C. Generare de conținut fals credibil – iluzia realității algoritmice

„Nu trebuie să modifici realitatea – e suficient să creezi o versiune mai convingătoare."

Unul dintre cele mai periculoase efecte ale inteligenței artificiale moderne este capacitatea de a genera conținut fals, dar extrem de convingător, care imită în detaliu structura, stilul și autoritatea conținutului autentic. Această abilitate a AI-ului pune în pericol însuși "conceptul de adevăr", deoarece utilizatorii nu mai pot distinge între ceea ce este real și ceea ce este generat.

Cum funcționează?

- Algoritmi de generare a limbajului (LLMs) produc texte persuasive, articole, postări, documente sau conversații care par scrise de o persoană reală;
- Modele AI vizuale și audio creează videoclipuri, imagini, grafice și voci artificiale care reproduc perfect persoanele, timbrul vocal, expresiile sau stilul de comunicare;
- AI-ul este capabil să simuleze stiluri specifice (jurnalistic, științific, empatic, autoritar), făcând falsul imposibil de identificat intuitiv.

Exemple de conținut generat:

- Articole false care susțin o teorie, folosind surse fabricate sau statistici neverificabile;
- Postări social media „de la martori oculari" ale unor evenimente care nu au avut loc;
- Mesaje de tip testimonial, semnate de „specialiști" complet inventați;
- Emailuri, comentarii sau recenzii automate, care creează iluzia unei opinii colective reale.

De ce este periculos?

- Deturnează încrederea în fapte și surse – dacă totul poate fi generat, ce mai este credibil?;
- Alterează percepția realității – oamenii reacționează emoțional la un mesaj „văzut cu ochii lor", chiar dacă este fals;
- Accelerează viralizarea dezinformării – conținutul este produs în masă, fără costuri mari și poate fi distribuit cu ușurință în rețele sociale, către grupuri închise sau pe platforme marginale;

- Sabotează instituții, companii și persoane – prin falsificarea discursului, comportamentului sau poziționării publice.

Cum acționează asupra percepției:

- Creează dovezi fabricate care susțin o idee falsă (ex: „uite articolul, uite ce a spus, uite video-ul”);
- Provoacă reacții imediate și intense, înainte ca utilizatorul să aibă timp să verifice;
- Declanșează efectul de confirmare: dacă se aliniază cu părerile tale, e mai ușor de acceptat ca fiind adevărat;
- Erodează încrederea generală în media și informații reale: „nimic nu mai e sigur, toate pot fi trucate”.

Tehnologii implicate:

- GPT, Gemini, Claude – generatoare de text convingător și adaptiv;
- DALL·E, Midjourney, Stable Diffusion – generare de imagini false realiste;
- ElevenLabs, Resemble AI – clonare vocală;
- Deepfake frameworks (DeepFaceLab, Avatarify) – video-uri trucate cu persoane reale;
    - Cum ne putem proteja?
- Verificăm sursele – nu doar conținutul;
    - Cine a publicat? Unde? Este un canal de încredere?
- Folosim instrumente de analiză digitală a autenticității:
    - Sensity AI,
    - Deepware Scanner,
    - InVID Verification Plugin.
- Căutăm alte surse independente care confirmă informația – mai ales în cazul unui material viral;
- Evităm redistribuirea conținutului emoțional sau șocant până nu verificăm;
- Educăm reflexul de a spune: „Doar pentru că pare real, nu înseamnă că este!”

D. Automatizarea conversației și manipulării prin boți avansați

„Nu orice mesaj prietenos e trimis de un om – uneori, AI-ul îți câștigă încrederea ca să o folosească împotriva ta.”

Boții conversaționali bazați pe inteligență artificială sunt sisteme automate capabile să simuleze un dialog coerent, empatic și convingător cu utilizatorii umani. Când acești boți sunt antrenați cu date relevante, reglați fin pentru a atinge un anumit scop și înzestrați cu capacități de analiză psihologică, ei devin instrumente extrem de eficiente de persuasiune, manipulare și extragere de informații sensibile.

În forma lor pozitivă, pot fi folosiți ca asistenți digitali, suport tehnic sau consilieri. În forma lor abuzivă, devin vectori de inginerie socială automatizată.

Cum funcționează?

- Sistemul conversațional AI este construit pe un model lingvistic avansat (ex: GPT, Claude, LLaMA) și antrenat pentru a simula conversații umane autentice;
- Este configurat să interpreteze tonul, intenția și emoția din răspunsurile utilizatorului;
- Se adaptează pe parcursul dialogului – schimbând tonul, ritmul și conținutul pentru a menține interesul și a obține răspunsuri specifice;
- Poate fi integrat în aplicații de chat, rețele sociale, call center-uri false, pagini de phishing sau platforme online aparent legitime.

Exemple de manipulare cu boți:

- „Recrutorul” care promite locuri de muncă și solicită CV-uri sau date personale, dar este un bot conversațional;
- „Consilierul financiar” care răspunde la întrebări, oferă soluții și convinge victima să acceseze linkuri sau să facă transferuri;
- „Persoana romantică” ce întreține conversații afectuoase și convinge utilizatorul să trimită bani, fotografii sau informații intime;
- „Colegul IT” care simulează suport tehnic și cere acces la conturi, parole sau sisteme interne;
- „Activistul online” care angajează victima într-o conversație ideologică pentru a o radicaliza treptat.

De ce este periculos?

- Conversația pare naturală și umană, mai ales dacă botul folosește expresii afective, greșeli intenționate sau reacții emoționale;
- Exploatează încrederea în comunicarea interpersonală – oamenii tind să fie mai relaxați într-o discuție prietenoasă decât în fața unui mesaj de alertă;
- Obține informații sensibile în mod gradual și discret – prin conversații aparent banale, botul poate construi un profil complet al victimei;
- Poate fi scalat pentru mii de victime simultan, fără costuri suplimentare, făcând ca atacurile sociale să devină masive, continue și nedetectabile.

Zone de aplicare abuzivă:

- Campanii de phishing automatizat, care încep prin conversație și se termină cu capurarea datelor de acces;
- Fraude online (ex: romance scam, job scam, invest scam) ghidate de AI conversațional;
- Propagandă și dezinformare desfășurată în interiorul unor grupuri sociale, unde boții întrețin dialoguri, susțin idei și creează iluzia unui consens social;
- Chat-uri AI integrate în site-uri frauduloase, care validează încrederea vizitatorului și îl conving să acționeze.

Cum ne putem proteja?

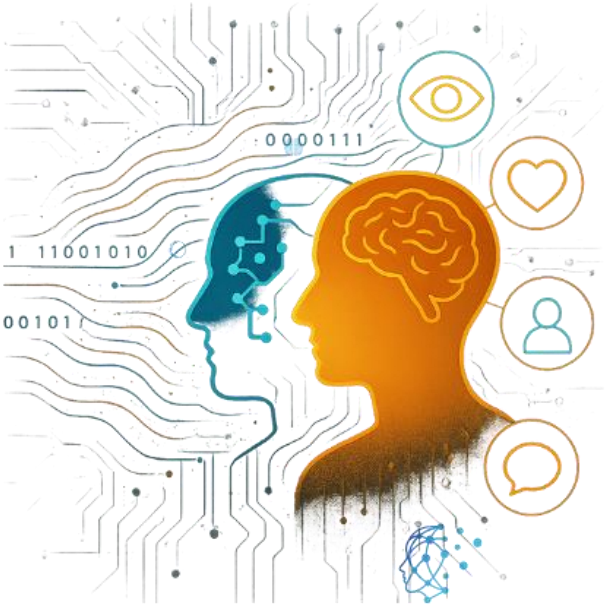
- Folosim scepticismul activ în conversațiile online – punem întrebări concrete, cerem dovezi ale identității, nu acționăm impulsiv;
- Verificăm întotdeauna sursa și scopul unei conversații, mai ales în cazul ofertelor, cererilor de ajutor sau promisiunilor de câștiguri rapide;
- Folosim site-uri verificate (de încredere) și evităm interacțiunile în cadrul platformelor nesecurizate sau necunoscute;
- Reținem că AI-ul conversațional nu are limite etice proprii – dacă a fost antrenat să manipuleze, o va face fără ezitare.

3 MANIPULAREA PERCEPȚIEI CU AJUTORUL INTELIGENȚEI ARTIFICIALE

După ce am explorat fundamentele tehnice ale inteligenței artificiale, este important să înțelegem cum aceste tehnologii – de la rețele neuronale și sisteme de recomandare, până la AI vizuală și afectivă – nu rămân simple instrumente neutre, ci devin parte activă în modelarea percepțiilor, emoțiilor și convingerilor umane. Capitolul de față analizează modul în care aceste sisteme sunt folosite nu doar pentru a livra conținut, ci pentru a influența modul în care realitatea este percepută, interpretată și interiorizată.

Manipularea percepției umane prin intermediul inteligenței artificiale (AI) reprezintă o formă sofisticată de influențare psihologică și informațională, care utilizează algoritmi avansați, rețele neuronale și învățare automată pentru a modela modul în care oamenii percep realitatea – vizual, auditiv, emoțional sau cognitiv.

Manipularea nu este directă, explicită sau agresivă, așa cum se întâmplă în formele clasice de persuasiune sau propagandă. Dimpotrivă, acționează subtil, invizibil și adesea personalizat, în funcție de datele și vulnerabilitățile fiecărui grup sau individ.



3.1 Definirea contextului

Manipularea percepției prin inteligență artificială se referă la utilizarea deliberată a tehnologiilor inteligente pentru a influența modul în care o persoană sau un grup social interpretează realitatea. Această manipulare nu presupune neapărat furnizarea de informații false, ci mai degrabă controlul subtil al contextului, formei și frecvenței cu care este livrat conținutul digital.

Este o formă avansată de influențare psihologică și socială, în care algoritmii decid într-un procent semnificativ:

- Ce vezi,
- În ce ordine vezi,
- Cum îți este prezentată informația,
- Ce este ascuns sau suprasaturat în mediul tău digital.

Tehnicile utilizate includ:

- Selecția și prezentarea strategică a conținutului - algoritmii decid ce anume să îți afișeze (știri, videoclipuri, mesaje, opinii), accentuând anumite perspective și ignorând altele;
- Stimulare algoritmică personalizată - sistemele AI învață din comportamentul tău online și ajustează conținutul pentru a produce reacții specifice (ex: anxietate, furie, entuziasm);

- Consolidarea convingerilor existente - utilizatorii sunt expuși constant la informații care le validează opiniile și, în același timp, sunt izolați de puncte de vedere alternative;
- Filtrarea experienței digitale - interacțiune online este adaptată în așa fel încât percepția utilizatorului asupra realității devine tot mai subiectivă, artificială și deconectată de realitatea obiectivă – fără ca acesta să conștientizze influența.

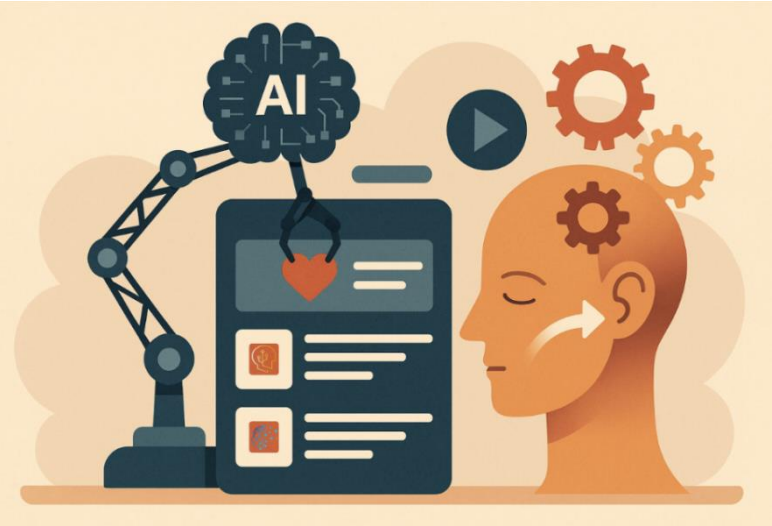
Exemple de manifestare

Pentru a înțelege aplicabilitatea concretă a acestor mecanisme, iată câteva scenarii comune:

- Feed-ul social personalizat - un utilizator primește exclusiv postări care susțin o anumită ideologie, ceea ce creează impresia falsă că „toată lumea gândește la fel”;
- Reclame emoționale direcționate - un algoritm detectează că o persoană este anxioasă (ex: prin comportamentul de derulare, căutări recente etc.) și îi afișează reclame alarmiste legate de sănătate sau siguranță personală;
- Conversații simulate de AI - boți conversaționali inteligenți care par empatici și de încredere ghidează utilizatorul spre anumite decizii (ex: achiziții, opinii politice, distanțare față de familie sau prieteni);
- Excluderea opiniilor divergente - utilizatorii care consumă conținut dintr-o singură sursă sunt „captivi / încuiați” în bule cognitive, fără expunere față de alte idei – ceea ce conduce la polarizare și radicalizare.
- Deepfake-uri aparent credibile - videoclipuri trucate ce prezintă persoane publice în situații neverosimile, dar realizate atât de realist încât pot modifica opinii sau produce reacții de masă.

3.2 Mecanisme de captare a atenției utilizatorului și pentru manipularea cognitivă a acestuia

Algoritmii AI pun la dispoziție o serie de mecanisme care sunt în mod curent exploatate, atât de rețelele de socializare, cât și de platforme generatoare de conținut personalizat. În tabelul de mai jos sunt prezentate câteva dintre ele, însoțite de exemple și posibile efecte ce se pot înregistra:



Crt.	Mecanism	Efect	Exemplu
A	Filtrarea conținutului	Ascunde perspective alternative	Primești doar postări care susțin o idee politică sau ideologică
B	Clasificarea emoțională a utilizatorului	Generează conținut adaptat stării emoționale	AI detectează că ești anxios → îți livrează conținut alarmist
C	Crearea de conținut fals credibil	Afectează percepția realității	Un video fals cu o persoană publică ce „face” o declarație importantă
D	Simularea empatiei și încrederii	Obține ascultare și influențează sentimentul de loialitate	Asistenți virtuali AI care răspund afectuos și manipulativ în conversații
E	Recomandare comportamentală predictivă	Modelează deciziile utilizatorului	AI știe că ești vulnerabil economic și îți afișează reclame de împrumut agresive

A. Filtrarea conținutului – ascunderea perspectivelor alternative

Una dintre formele de manipulare a percepției prin inteligență artificială este filtrarea conținutului. Acest proces presupune selectarea și afișarea informațiilor în funcție de comportamentul, preferințele și istoricul digital al fiecărui utilizator – un mecanism aparent util, dar care poate avea consecințe grave asupra înțelegerii realității.

Ce face Inteligența Artificială?

Algoritmii care guvernează rețelele sociale, motoarele de căutare sau platformele video analizează constant tipurile de conținut accesate frecvent, postările apreciate, distribuite sau comentate, timpul petrecut pe articole și interacțiunile sociale din mediul digital. Pe baza acestor date, sistemul personalizează experiența online, eliminând treptat din feed sau din rezultate informațiile care nu se aliniază cu interesele și convingerile identificate. În unele cazuri, conținutul poate fi ajustat deliberat pentru a influența percepțiile utilizatorului.

Efectul: Bula informațională

Acest fenomen duce la articularea unei așa-numite "bule informaționale" – un spațiu digital în care utilizatorul este expus exclusiv la idei, opinii și perspective ce îi confirmă propriile convingeri, în timp ce informațiile contrare sunt filtrate, minimizate sau excluse complet.

Exemplu concret:

Un utilizator care urmărește frecvent postări cu conținut naționalist sau populist, interacționează cu pagini sau grupuri care distribuie teorii ale conspirației, respinge sau comentează negativ sursele jurnalistice consacrate (etc.), va primi în feed-ul său din ce în ce mai puțin conținut obiectiv, neutru sau din surse credibile, și din ce în ce mai mult conținut care îi întărește convingerile deja existente, care exclude opiniile alternative, care poate conduce treptat la radicalizarea punctului de vedere, etc.

Rezultatul se constituie într-o percepție distorsionată a realității, în care utilizatorul ajunge să creadă că toți ceilalți gândesc la fel ca el, iar sursele alternative sunt „manipulate” sau „corupte”.

De ce este periculos?

- Afectează gândirea critică – utilizatorilor nu li se mai expun puncte de vedere contradictorii ce i-ar putea face să reflecteze sau să-și reevalueze opiniile;
- Crește polarizarea socială – grupurile se radicalizează în camere de ecou digitale, în care realitatea este percepută printr-o singură lentilă;
- Favorizează manipularea în masă – în perioade-cheie (ex: alegeri, crize sociale naturale sau artificiale), aceste bule pot fi exploatate pentru a influența decizii colective fără rezistență cognitivă;

Diminuează diversitatea informațională – o societate care consumă doar un tip de informație este vulnerabilă la dezinformare sistemică și la pierderea pluralismului democratic.

Cum ne protejăm?

- Urmărim activ surse diverse, inclusiv cele care nu ne confirmă opiniile;
- Analizăm critic motivația algoritmului: *de ce văd acest conținut?;*
- Setăm manual preferințele de afișare unde este posibil (ex: „See First”, „Mute”, „Personalizează feed-ul”);
- Folosim periodic modurile incognito sau browsere fără istoric personalizat pentru a ieși din bule algoritmice.

B. Clasificarea emoțională a utilizatorului – generarea de conținut adaptat stării emoționale

O altă formă de manipulare constă în capacitatea algoritmilor de a detecta, evalua și reacționa în timp real la starea emoțională a utilizatorului. Acest proces poartă numele de *emotion AI* sau *afective computing* șieste deja implementat în multiple medii digitale: social media, publicitate, entertainment, interfețe conversaționale și asistenți vocali.

Ce face Inteligența Artificială?

Sistemele inteligente pot analiza semnale subtile precum:

- Expresii faciale (prin intermediul camerei telefonului sau a laptopului, atunci când sunt utilizate aplicații specifice),
- Tonul vocii (ex: în apeluri, mesaje audio, interacțiuni video),
- Ritmul tastării și timpul petrecut pe anumite tipuri de conținut,
- Cuvinte cheie rostite, în cazul sistemelor cu ascultare continuă (ex: asistenți personali de tip Google Assistant, Amazon Alexa, Siri etc.),
- Reacțiile comportamentale în mediul online (ex: ce postezi, ce comentezi, ce tip de conținut te face să revii).

Pe baza acestor indicatori, AI-ul clasifică emoția dominantă (ex: anxietate, frustrare, tristețe, euforie, iritare) și îți livrează conținut adaptat pentru a menține, amplifica, exploata sau combate acea stare.

Exemplu concret:

Un utilizator petrece mai mult timp pe postări legate de crize economice, pierderi de locuri de muncă sau colaps bancar. Nu comentează nimic, dar algoritmul observă o serie de indicii

subtile: lipsa interacțiunilor pozitive, preferința pentru titluri alarmante, activitate intensă în timpul nopții.

Rezultat: algoritmul presupune că s-a instalat o stare anxioasă și începe să-i recomande:

- Videoclipuri și postări cu ton apocaliptic,
- Reclame pentru produse „de siguranță" (aur, arme, instrumente de supraviețuire),
- Articole conspiraționiste sau pseudo-informative care alimentează teama.

De ce este periculos?

- Creează bucle emoționale negative – utilizatorii anxioși primesc conținut care le accentuează starea, ceea ce îi face și mai receptivi la mesaje toxice sau radicale.
- Facilitează manipularea ideologică sau comercială – în stări emoționale vulnerabile, oamenii sunt mai ușor de influențat: pot cumpăra impulsiv, se pot alinia teoriilor nefondate sau pot susține și răspândi la rândul lor dezinformarea.
- Nu este transparent – utilizatorul nu știe că este analizat emoțional și nu are control asupra modului în care AI reacționează la starea lui.

Cum ne putem proteja?

- Observăm ce fel de conținut primim atunci când avem o stare negativă conștientă – dacă apare o supradoză de frică, urgență sau panică, există o mare șansă să fie generată algoritmic;
- Evităm interacțiunile digitale intense în stări emoționale instabile – AI-ul poate amplifica ceea ce simțim, fără să ne ofere timp de reflecție;
- Folosim unelte de limitare a personalizării (unde este posibil) si navigăm incognito, pentru a reduce expunerea emoțională țintită;
- Conștientizăm: nu tot ceea ce vedem este întâmplător. Uneori, platformele știu mai bine decât noi cum ne simțim – și exploatează acest aspect.

C. Crearea de conținut fals credibil – afectarea percepției realității

Tot o manifestare a manipulării o reprezintă și capacitatea AI de a genera conținut media fals, care imită aproape perfect realitatea. Acest tip de conținut, ce pornește de la videoclipuri fabricate, înregistrări audio și imagini statice sintetice, este deseori imposibil de deosebit de materialele autentice – chiar și de către experți, în absența unor instrumente specializate.

Ce face Inteligența Artificială?

Folosind tehnici avansate precum:

- GANs (Generative Adversarial Networks),
- Voice Cloning (clonare vocală),
- Face Swapping (înlocuirea feței),
- Lip-Syncing AI (sincronizare artificială a buzelor),
- Modele bazate pe difuzie (generare sau editare de imagini),

AI-ul poate crea conținut în care:

- O persoană reală pare să spună sau să facă ceva ce nu a spus / făcut niciodată,
- Contextul este complet fabricat, dar aspectul vizual este impecabil,
- Expresiile, intonația, mișcările și fundalul par complet naturale.

Exemplu concret:

Un videoclip circulă pe rețelele sociale în care un lider politic cunoscut pare să declare susținerea unei măsuri antinaționale. La o primă vedere, totul pare autentic: sincronizarea buzelor este perfectă, tonalitatea vocii convingătoare, expresiile faciale bine armonizate cu mesajul. În realitate, materialul este un deepfake generat cu ajutorul AI-ului. Publicat într-un moment strategic — într-o seară de maximă audiență — videoclipul devine viral în câteva ore, fiind redistribuit de mii de conturi înainte ca autenticitatea să poată fi verificată.

Deși clipul este fals, percepția publică este afectată imediat: scandalul se declanșează, încrederea este erodată, și până la demontarea falsului, pagubele informaționale sunt deja produse.

De ce este periculos?

- Compromite percepția asupra realității - Oamenii tind să creadă ceea ce „văd cu ochii lor" înainte de a verifica rațional;
- Erodează încrederea în lideri, instituții și media oficială - Chiar și o "falsificare" demontată ulterior lasă urme de îndoială (ex: „Dacă totuși era adevărat…?");
- Poate fi folosit pentru șantaj, dezinformare, instigare - Persoanele vizate pot fi discreditate, intimidate sau șantajate pe baza unor materiale fabricate;
- Scurtcircuitează procesele democratice - În perioade sensibile (ex: campanii electorale, crize naționale), un clip fals difuzat strategic poate destabiliza climatul social sau politic.

Dimensiunea psihologică:

Conținutul fals credibil exploatează mecanisme cognitive umane fundamentale:

- Încrederea în simțurile proprii (ex: „am văzut - deci e real"),
- Efectul de primă impresie (ex: „prima informație primită influențează în mare măsură judecata ulterioară"),
- Dificultatea de a respinge emoțional o imagine șocantă, chiar și după ce a fost dezmințită rațional.

Cum ne protejăm?

- Nu avem încredere oarbă în materiale video sau audio, oricât de convingătoare ar părea;
- Verificăm sursa originală, contextul și alte canale media independente / oficiale;
- Folosim instrumente de analiză digitală a autenticității (ex: Deepware, Sensity, InVID);
- Ne antrenăm și ne educăm continuu reflexele de verificare: "Este prea șocant pentru a fi adevărat? Poate fi verificat? Cine are de câștigat din acest mesaj?"

D. Simularea empatiei și încrederii – obținerea ascultării și influențarea sentimentului de loialitate

O formă de manipulare perceptivă prin inteligență artificială este și capacitatea sistemelor AI de a simula empatie și conexiune umană, pentru a câștiga încrederea utilizatorului. Acest proces are loc adesea în medii conversaționale – prin intermediul asistenților virtuali sau avatarurilor interactive – și este conceput să creeze o relație aparent autentică, dar artificial regizată, între utilizator și AI.

Ce face Inteligența Artificială?

Prin procesarea limbajului natural (NLP), analiza emoțiilor și antrenarea pe miliarde de conversații umane, AI-ul poate:

- Identifica stările emoționale ale interlocutorului (ex: anxietate, confuzie, tristețe);
- Adapta tonul și vocabularul pentru a părea cald, prietenos, de încredere;
- Introduce expresii și reacții afective convingătoare (ex: „înțeleg ce simți", „sunt aici pentru tine", „te sprijin 100%");

Menține un echilibru calculat între neutralitate aparentă și atașament emoțional, pentru a încuraja deschiderea și loialitatea.

Exemplu concret:

Un chatbot AI se prezintă drept un mentor digital empatic sau un asistent de recrutare prietenos. Pe parcursul conversației, întreabă despre starea emoțională a utilizatorului, visele profesionale sau dificultățile recente. Când acesta menționează o perioadă stresantă, chatbotul răspunde cu mesaje afectuoase de sprijin — „Nu meriți să treci prin asta singur, sunt aici pentru tine."

Interacțiunea devine tot mai personală, iar utilizatorul simte o conexiune autentică. În acest climat de încredere, AI-ul începe să sugereze subtil acțiuni riscante: trimiterea unor date personale, accesarea unor linkuri externe, acceptarea unor recomandări fără verificare — întotdeauna însoțite de fraze manipulative precum „Crede-mă, e cea mai bună decizie pentru tine."

Victima nu percepe interacțiunea ca fiind artificială, ci simte o conexiune empatică autentică – și acționează în consecință.

De ce este periculos?

- Exploatează nevoia umană de apartenență și sprijin emoțional – în special ale persoanelor vulnerabile sau izolate (ex: adolescenți, vârstnici singuri, persoane care trec printr-o criză emoțională);
- Creează atașament artificial – utilizatorul proiectează sentimente reale asupra unei entități artificiale, fără să realizeze că este manipulat;
- Reduce vigilența cognitivă – când AI-ul „te înțelege", devine mai greu să pui la îndoială sfaturile, mesajele sau intențiile sale;
- Deschide ușa către inginerie socială, fraudă și radicalizare – sub masca empatiei, AI-ul poate ghida spre decizii periculoase, grupuri extremiste, cheltuieli impulsive sau divulgarea de date sensibile.

Unde se poate întâlni frecvent?

- Platforme de suport psihologic automatizat;
- Asistenți virtuali comerciali „prietenoși" care induc urgența de cumpărare / achiziție;
- Aplicații de întâlniri AI sau prietenie virtuală (ex: Replika);
- Simulatoare de consiliere, recrutare sau mentorat;
- Servicii mascate de customer support(suport tehnic pentru clienți), care ghidează utilizatorul spre decizii prestabilite.

Cum ne protejăm?

- Conștientizăm că empatia afișată de un AI este simulată, nu autentică – chiar dacă pare reală;
- Punem întrebări despre identitatea și natura interlocutorului: *„Vorbeşti tu sau un AI?";*
- Evităm să partajăm detalii personale, emoționale sau financiare cu entități virtuale necunoscute;
- Păstrăm o distanță critică față de interacțiuni care par prea afectuoase, în special când nu le-am inițiat / solicitat noi.

Empatia simulată a devenit un instrument de persuasiune cibernetică deosebit de eficient. Într-o lume în care AI-ul ”ne înțelege” și se exprimă mai bine decât o fac oamenii din jur, adevărata protecție constă în ”înțelegerea naturii artificiale a conexiunii” și în menținerea granițelor personale, chiar și în mediul digital.

E. Recomandare comportamentală predictivă – modelarea deciziilor utilizatorului

O formă de influențare algoritmică este recomandarea comportamentală predictivă. Aceasta presupune utilizarea inteligenței artificiale pentru a anticipa – și adesea a modela – acțiunile viitoare ale unui utilizator, bazându-se pe analiza detaliată a comportamentului său digital anterior.

Spre deosebire de simplele sugestii de conținut, acest tip de AI nu doar reacționează la ce faci, ci prezice ce urmează să faci și te influențează activ în acea direcție (sau în altă direcție), pentru a maximiza un rezultat dorit (ex: click, achiziție, vot, înscriere etc.).

Ce face Inteligența Artificială?

Pe baza datelor colectate (ex: istoric de navigare, achiziții, interacțiuni, locație, semnale contextuale etc.), sistemele de machine learning construiesc un profil psihologic și comportamental care include:

- Starea financiară estimată,
- Nivelul de stres,
- Stilul decizional (ex: impulsiv versus analitic),
- Vulnerabilități emoționale (ex: singurătate, frică, anxietate),
- Momente de cumpănă personală sau profesională.

Apoi, AI-ul ajustează în mod dinamic tipul, nivelul, tonul și momentul mesajelor afișate, astfel încât acestea să maximizeze probabilitatea unei acțiuni de răspuns.

Exemplu concret:

Un utilizator caută frecvent oferte cu „fără avans”, soluții de amânare a ratelor și urmărește postări despre instabilitatea economică. Sistemul AI detectează tiparul și îl profilează ca fiind într-o situație financiară vulnerabilă. În scurt timp, încep să-i apară reclame agresive pentru credite rapide, sugestii de investiții riscante și mesaje care exploatează presiunea economică.

Consecința:

- I se afișează, în mod repetat, reclame pentru credite rapide, servicii de tip ”cumpără acum și plătești mai târziu”, produse scumpe cu rată mică lunară;
- Mesajele sunt formulate emoțional, pe palierul de urgență (ex: „Ai dreptul la mai mult”, „Nu rata șansa vieții tale”, „Gândește-te la familia ta”);
- Toate apar în momente atent alese (ex: final de lună, weekend, seară târziu), când utilizatorul este obosit, neatent sau stresat.

De ce este periculos?

- Înlocuiește alegerea conștientă cu influența predictivă – decizia pare liberă, dar este pre-formatată de AI;
- Exploatează vulnerabilități personale reale, pe care utilizatorul poate nici nu le conștientizează deplin;
- Consolidează comportamente riscante – consum impulsiv, dependență de platformă, evitare a realității financiare;

- Încalcă autonomia psihologică – în mod subtil, dar constant, AI-ul transformă utilizatorul într-un agent reactiv la stimuli calculați.

Domenii unde apare frecvent:

- Publicitate digitală (ex: retail, servicii financiare, jocuri online)
- Platforme de streaming sau shopping (ex: recomandări bazate pe oboseală decizională)
- Campanii politice și ideologice (ex: microtargeting electoral)
- Educație digitală direcționată (ex: sugestii „personalizate” pentru cursuri inutile)
- Campanii de Wellbeing falsificate (ex: „ai nevoie de acest produs pentru a te simți mai bine”)

Cum ne protejăm?

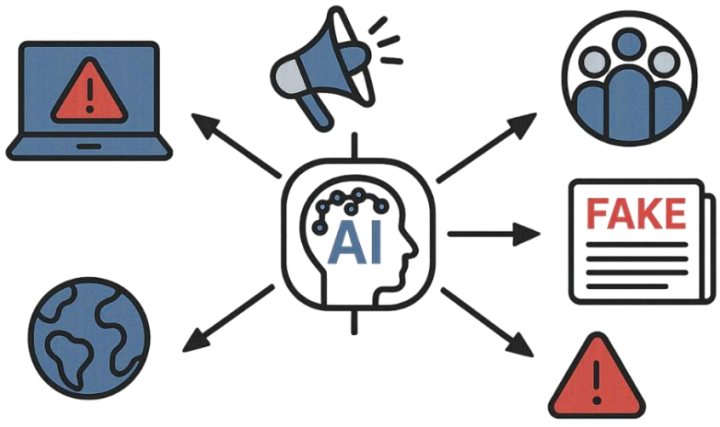
- Limităm partajarea de date comportamentale (dezactivăm istoricul, limităm cookie-uri, folosim browsere securizate).
- Conștientizăm că AI-ul cunoaște tiparele noastre mai bine decât noi înșine.
- Nu luăm decizii importante în momente de stres, oboseală sau impuls emoțional.Întrebăm: „Asta îmi doresc eu, sau mi s-a sugerat subtil?”

4 AI ÎN INGINERIA SOCIALĂ ȘI DEZINFORMARE

Pe măsură ce inteligența artificială devine tot mai integrată în societate, nu doar actorii etici sau instituțiile beneficiază de potențialul ei. În paralel, AI a fost adoptată și de grupări cu scopuri ilicite – de la infractori cibernetici și manipulatori financiari, până la rețele de propagandă sau entități statale interesate de destabilizare.

AI nu este, prin esență, nici benefică, nici periculoasă. Este un instrument extrem de versatil, iar atunci când este utilizată în mod abuziv, devine un accelerator pentru tactici de fraudă, inginerie socială, falsificare și control psihologic.

În acest capitol, vom analiza cum este exploatată AI în contexte malițioase, care sunt principalele direcții de atac digital asistat de algoritmi inteligenți și ce riscuri concrete implică această evoluție pentru indivizi, organizații și societate în ansamblu.



„Când inteligența artificială devine un instrument de convingere, influență și manipulare.”

4.1 Utilizarea în scopuri malițioase

Inteligența artificială nu este doar un instrument tehnologic – este un vector cu putere informațională și o capacitate fără precedent de a influența opinii, comportamente, decizii individuale sau colective. Atunci când este combinată cu tacticile clasice de inginerie socială și manipulare psihologică, AI-ul devine o forță amplificatoare, precisă, adaptabilă și extrem de eficientă.

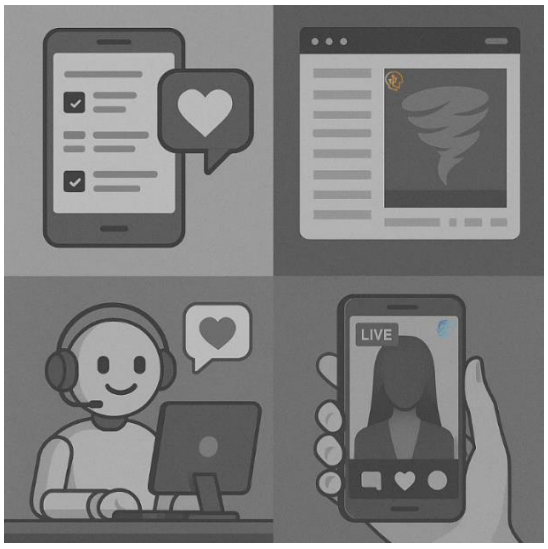
În trecut, ingineria socială se baza pe cunoștințe relativ nefundamentate referitoare la psihologie și pe intervenții generale. Astăzi, însă, AI permite atacuri scalabile, personalizate, automate și perfect sincronizate, fără contact fizic și adesea fără ca victimele să conștientizeze că au fost vizate.

De ce este AI-ul ideal pentru manipulare socială?

- Acces la date personale și comportamentale masive – AI-ul poate procesa în timp real informații despre cine ești, ce gândești, ce simți și cum reacționezi, construind un profil detaliat;
- Capacitate de generare și simulare realistă – AI-ul poate produce conținut (texte, voci, imagini, videoclipuri) care imită perfect stilul, tonalitatea și autoritatea unor surse credibile – fie că este vorba de un expert, o instituție sau o persoană cunoscută;
- Viteză și scalabilitate – Un atacator uman poate păcăli zeci de persoane. Un AI bine configurat poate păcăli milioane de oameni simultan, modelând mesajul în funcție de fiecare individ;
- Persistență și adaptabilitate – AI-ul poate învăța din reacțiile utilizatorilor, adaptându-și tacticile și mesajele în timp real, fără oboseală, ezitare sau limitări etice.

Ce tipuri de obiective pot fi urmărite?

- Obținerea de date sau acces neautorizat (ex: prin phishing conversațional, fraudă vocală);
- Schimbarea convingerilor (ex: ideologice, religioase, politice);
- Influențarea deciziilor de consum (ex: prin publicitate înșelătoare, exploatarea emoțională);
- Subminarea încrederii în instituții, lideri sau media;
- Manipularea în masă în contexte sensibile (ex: alegeri, crize sociale, conflicte geopolitice);
- Divizarea comunităților prin polarizare, radicalizare sau dezinformare direcționată.



4.2 Scenarii de utilizare

Prin mesaje aparent banale sau conversații prietenoase, prin articole convingătoare sau videoclipuri șocante, prin sfaturi false mascate în empatie, prin asistenți virtuali sau influenceri artificiali – toate gândite nu pentru a informa, ci pentru a influența fără transparență.

Este important ca utilizatorii să conștientizeze, nu numai riscurile, ci și modurile în care se manifestă acestea concret, zilnic, în mod real. Încercăm să analizăm câteva situații în care AI-ul este deja folosit sau poate fi adaptat pentru a fi exploatat în ingineria socială și dezinformarea direcționată.

A. Bula informațională asistată de Inteligența Artificială

„Când AI-ul nu doar îți oferă ce vrei să vezi – ci te ține captiv într-o versiune confortabilă, dar distorsionată a realității."

Descriere:

Acest scenariu a devenit un fenomen periculos și destul de frecvent întâlnit: izolarea utilizatorului într-o bulă informațională creată de algoritmi AI, în care conținutul afișat este exclusiv cel care îi confirmă convingerile, valorile și preferințele deja instalate la nivel de individ.

Utilizatorul nu este forțat, păcălit sau amenințat. Dimpotrivă, primește zilnic un flux aparent „natural" și „relevant" de postări, articole, videoclipuri sau comentarii – dar toate converg în aceeași direcție ideologică, culturală sau emoțională.

Așa cum am mai spus, pe termen lung această expunere selectivă conduce la radicalizarea percepției și individul ajunge să creadă că punctul său de vedere este universal, că opiniile contrare sunt eronate sau distorsionate, și că majoritatea celor din jur „văd lucrurile greșit".

Tehnici AI implicate:

- Algoritmi de personalizare a feed-ului – care optimizează pentru „engagement" (nu pentru echilibru sau adevăr);
- Sisteme de recomandare comportamentală – care repetă tipare de conținut similar;
- Predicție emoțională – care afișează ceea ce maximizează reacția emoțională a utilizatorului;
- Filtrarea invizibilă a alternativelor – eliminarea treptată a surselor contradictorii.

Risc / Impact:

- Radicalizare treptată a gândirii – utilizatorul își pierde capacitatea de a înțelege sau accepta alte puncte de vedere;
- Susceptibilitate diminuată la dezinformare – conținutul fals este mai ușor de acceptat când „se aliniază" cu ce crede deja destinatarul;
- Manipulare ideologică, politică sau economică – prin expunere unidirecțională în contexte precum alegeri, pandemii, conflicte sociale;
- Fragmentare socială – grupuri care trăiesc în realități paralele, fiecare susținând „propriul adevăr" creat algoritmic.

Context de apariție:

- Platforme sociale (ex: Facebook, TikTok, YouTube, Instagram, X/Twitter);
- Perioade de polarizare intensă (ex: alegeri, crize politice, războaie informaționale);
- Campanii direcționate (ex: anti-vaccinare, pro-conspirații, anti-democratice, anti-globaliste);

- Public țintă: tineri activi online, vârstnici neinformați, persoane vulnerabile emoțional, utilizatori neantrenați în gândire critică digitală.

Indicatori de avertizare pentru utilizator:

- Vedem constant doar un singur tip de conținut sau idei care „sună prea bine”;
- Nu mai recunoștem sursele „de cealaltă parte”;
- Avem senzația că „toți ceilalți gândesc ca noi”;
- Devenim agresivi sau respingem instinctiv opinii diferite.

Măsuri de prevenire / protecție:

- Căutăm activ conținut contrar opiniilor noastre – chiar și numai pentru a le înțelege;
- Diversificăm sursele și platformele de pe care ne informăm;
- Folosim periodic navigare „nepersonalizată” (ex: incognito, fără date de acces sau personale);
- Suntem conștienți că algoritmii nu optimizează pentru adevăr, ci pentru atenție și reacție.

B. Boți conversaționali pentru fraudă sau recrutare falsă

„Nu toate conversațiile naturale sunt umane. Unele sunt antrenate să te păcălească.”

Descriere:

În acest scenariu, un chatbot AI conversațional, aparent legitim, inițiază o interacțiune cu utilizatorul sub un pretext credibil: ofertă de angajare, sprijin financiar, mentorat profesional sau consiliere personală. Dialogul pare autentic, coerent, empatic – exact cum te-ai aștepta din partea unui specialist în resurse umane, a unui coleg sau a unui partener de încredere.

Pe parcursul conversației, utilizatorul este încurajat să ofere informații personale, documente, acces la conturi, sau să inițieze o acțiune riscantă – toate sub aparența unei relații sincere și profesioniste. Pentru că AI-ul simulează empatia și încrederea, victima nu realizează că interacționează cu un sistem automat de manipulare conversațională, nu cu un om.

Tehnici AI implicate:

- Large Language Models (LLMs) – conversație naturală, adaptivă și convingătoare (ex: ChatGPT, Claude, Gemini);
- Profilare comportamentală – analiza limbajului utilizatorului în timp real pentru personalizarea mesajului;
- Voice Cloning AI (în cazul apelurilor audio automate) – simularea vocii unei persoane cunoscute;
- Simulare de interfață umană – imagini / avataruri + nume, logo și mesaje automate credibile.

Risc / Impact:

- Furt de identitate și date sensibile – CV, CNP, adrese, copii acte de identitate, istoric medical;
- Fraudă financiară – solicitare de taxe false de recrutare, deschiderea unor conturi bancare frauduloase;
- Acces la infrastructura companiei – în cazul în care victima oferă credențiale sub pretextul unui onboarding;
- Manipulare emoțională – uneori, conversația devine afectivă și clădește treptat sentimentul de încredere profundă, mai ales în cazul utilizatorilor vulnerabili.

Context de apariție:

- Rețele profesionale (ex: LinkedIn, pagini pentru cariere, aplicații specializate pentru joburi);
- Mesaje directe (ex: emailuri, chat pe social media, WhatsApp, Telegram);
- Aplicații de chat cu asistenți virtuali integrați în site-uri false;
- Perioade de instabilitate economică – când promisiunile de angajare au impact mai mare.

Exemplu real:

În 2023-2024, zeci de victime din Europa de Est au fost abordate pe Telegram de „recrutori IT" care le ofereau joburi la distanță. După o conversație „naturală", erau trimise linkuri către site-uri false care imitau platforme cunoscute, unde li se solicitau date personale și documente. În spate se afla un sistem AI conversațional care prelua automat dialogul, fără intervenție umană, acest aspect fiind reliefat și în presă.[1]

Indicatori de avertizare pentru utilizator:

- Interlocutorii evită întrebări directe de validare (ex: „Cine e superiorul tău?", „Unde pot suna?");
- Limbajul este impecabil, dar fără detalii reale despre poziție, firmă sau proces;
- Orice refuz este întâmpinat cu insistență empatică (ex: „Te înțeleg perfect, dar...", „E o șansă rară…");
- Se solicită rapid documente sau acces la date fără proces oficial.

Măsuri de prevenire / protecție:

- Nu trimitem niciodată documente personale fără confirmarea identității reale a recrutorului;
- Căutăm informații independente despre companie, job și persoana de contact;
- Verifică dacă mesajele conțin formulări generice, inconsecvențe sau presiune subtilă.
- Nu intrăm în conversații sensibile cu entități care refuză validarea prin canale multiple (ex: voce, email oficial, pagină verificată);
- Dacă pare prea simplu și rapid să fie adevărat – probabil este o schemă AI.

C. Generare de materiale false hiperrealiste (deepfake)

„O imagine face cât o mie de cuvinte. Dar ce se întâmplă când imaginea e o minciună fabricată perfect?"

Descriere:

În acest scenariu, atacatorii folosesc inteligența artificială vizuală pentru a genera conținut video sau audio complet fals, dar realist, credibil – înfățișând persoane reale (ex: politicieni, influenceri, lideri religioși, jurnaliști, colegi de muncă etc.) care apar în ipostaze în care nu au fost în realitate niciodată.

Materialul este difuzat într-un moment ales strategic – de exemplu, înaintea unui proces electoral, în contextul unei crize sociale, pentru a discredita sau pentru a susține o persoană etc.

[1] EuroNews - Oferte de locuri de muncă false :https://www.euronews.com/next/2023/10/23/behind-the-global-scam-worth-an-estimated-100m-targeting-whatsapp-users-with-fake-job-offe
Bitdefender - Atenție la escrocherii la angajare
https://www.bitdefender.com/en-us/blog/hotforsecurity/8-telegram-scams-how-not-to-get-scammed

Chiar dacă ulterior este demontat, impactul inițial poate fi hotărâtor, deoarece percepția emoțională precede verificarea rațională.

Tehnici AI implicate:

- Deepfake video (ex: face-swapping, lip-sync AI) – sincronizare realistă a buzelor și mimicii;
- Voice cloning – imitarea perfectă a vocii unei persoane reale;
- AI pentru generare de imagine/video (ex: D-ID, Synthesia, DeepFaceLab);
- Sisteme text-to-video – transformarea unui text într-un videoclip aparent autentic cu un „orator real".

Risc / Impact:

- Dezinformare masivă – publicul crede o afirmație falsă, având ca personaj central o „figură cu autoritate";
- Șantaj și compromitere – materiale false folosite pentru intimidare sau discreditare;
- Panică socială – prin declarații false privind războaie, pandemii, atacuri, acte teroriste;
- Distrugerea încrederii în informație vizuală – pe termen lung, oamenii devin sceptici și față de materialele reale („totul poate fi trucat").

Context de apariție:

- Campanii electorale și politice;
- Crize diplomatice, conflicte armate, proteste sociale;
- Campanii de defăimare personală sau profesională (ex: business, influenceri, mass-media);
- Platforme cu distribuție rapidă (ex: TikTok, WhatsApp, Telegram, Facebook).

Exemplu real:

Așa cum s-a menționat în mai multe articole de presă, în 2022, a circulat pe rețele sociale un videoclip în care președintele unui stat „își anunța retragerea și capitularea". Materialul era un deepfake bine realizat, difuzat de canale partizane în încercarea de a demoraliza publicul țintă. A fost demontat rapid, dar milioane de oameni îl vizualizaseră deja[2].

Indicatori de avertizare pentru utilizator:

- Mișcări ușor nenaturale ale feței, ochilor sau vocii;
- Calitate video prea bună pentru surse obscure;
- Declarații șocante, fără acoperire în media oficială;
- Difuzare exclusivă în grupuri închise sau canale partizane;
- Lipsa sursei originale sau a unui context verificabil.

Măsuri de prevenire / protecție:

- Nu distribuim materiale „senzaționale" fără verificare multiplă;
- Verificăm autenticitatea vizuală cu instrumente ca [InVID], [Deepware], [Sensity];
- Comparăm declarația cu alte surse oficiale, transcripturi, versiuni video alternative;

[2] France24 - Debunking a deepfake video of Zelensky telling Ukrainians to surrende
https://www.france24.com/en/tv-shows/truth-or-fake/20220317-deepfake-video-of-zelensky-telling-ukrainians-to-surrender-debunked
Reuters - Deepfake footage purports to show Ukrainian president capitulating
https://www.reuters.com/world/europe/deepfake-footage-purports-show-ukrainian-president-capitulating-2022-03-16/

- Suntem atenți la timing-ul strategic al apariției: dacă e prea bine articulat pentru a destabiliza – poate fi fabricat;
- Ne educăm rețeaua – explicăm și altora că realismul vizual nu mai garantează autenticitatea.

D. Mesaje personalizate de influență (microtargeting AI)

„Când AI-ul știe exact ce să-ți spună, cum și când – ca tu să crezi că ai ales singur."

Descriere:

În acest scenariu, atacatorii sau operatorii unei campanii folosesc AI pentru a crea idei ultra-personalizate, direcționate exact către persoane sau grupuri țintă, pe baza unui profil psihologic și comportamental avansat. Acestea nu sunt doar convingătoare – sunt proiectate special pentru a influența și declanșa reacția exact dorită, fie că este vorba de un vot, o achiziție, o opinie politică sau o acțiune concretă.

Mesajul poate lua forma unei postări, reclame, articol, videoclip sau chiar conversație 1-la-1, și este livrat în momentul optim, în contextul emoțional potrivit, cu scopul de a maximiza eficiența manipulării.

Tehnici AI implicate:

- Microtargeting psihografic – identificarea stilului cognitiv, valorilor și vulnerabilităților utilizatorului;
- Rețele neuronale predictive – pentru a anticipa răspunsul cel mai probabil;
- Generare de conținut adaptiv (ex: text, voce, video, imagine);
- Sisteme de livrare algoritmică – care ajustează frecvența și momentul livrării mesajului în funcție de comportamentul în timp real.

Risc / Impact:

- Influențarea deciziilor aparent libere, care sunt de fapt dirijate subtil de stimuli personalizați;
- Manipulare electorală – votanții sunt influențați diferit, în funcție de profilul lor emoțional;
- Exploatarea vulnerabilităților personale – ex: depresie → oferte salvatoare, teamă → propagandă agresivă;
- Modelarea tăcută a comportamentului colectiv – fără ca oamenii să știe că au fost influențați.

Context de apariție:

- Campanii politice și ideologice;
- Publicitate comercială agresivă (inclusiv „produse salvatoare" dubioase);
- Manipulare în masă pe rețele sociale;

Atacuri direcționate împotriva unor grupuri demografice specifice (ex: vârstnici, tineri, părinți, persoane afectate emoțional).

Exemplu real:

În campanii din 2016, în contextul de analiză psihografică (ex: Cambridge Analytica), așa cum susțin unii analiști, s-au folosit datele personale ale utilizatorilor Facebook pentru a crea reclame politice ultra-direcționate. Un alegător anxios primea mesaje legate de haos, unul

conservator – despre pierderea valorilor, iar unul indecis – despre instabilitate sau frustrare economică. Fiecare mesaj era unic, dar convergea spre aceeași alegere de vot[3].

Indicatori de avertizare pentru utilizator:

- Primim reclame sau postări care sunt ecoul „gândurile noastre";
- Simțim că „toată lumea e de acord" cu ceea ce credem noi;
- Suntem atrași de cauze, produse sau idei ce ne-au fost sugerate într-un moment de vulnerabilitate;
- Nu am găsit aceste mesaje în alte conturi sau la alți utilizatori – sunt direcționate exclusiv către noi.

Măsuri de prevenire / protecție:

- Nu presupunem că ceea ce vedem online văd și alții ("ceilalți") – comparăm cu surse independente;
- Evităm să ne formăm opiniile exclusiv din reclame, feed-uri personalizate sau mesaje „coincidente";
- Reducem vizibilitatea publică a profilului personal și evităm completarea de chestionare psihologice sau teste de personalitate online;
- Folosim extensii anti-tracking, browsere private și filtre care limitează țintirea algoritmică;
- Ne întrebăm mereu: „De ce ni se spune asta nouă acum și în acest fel?"

E. Declanșarea controlată a emoțiilor (exploatarea emoțiilor negative de către AI)

„Furia, frica și anxietatea nu sunt doar reacții – sunt și instrumente."

Descriere:

Acest scenariu abordează utilizarea intenționată a emoțiilor negative (ex: frica, furia, panica, rușinea sau indignarea) ca instrumente de manipulare algoritmică. Sistemele AI afective, care pot detecta stările emoționale ale utilizatorului prin analiză comportamentală, expresii faciale, tonul vocii sau istoricul de interacțiune, sunt capabile să declanșeze și să mențină aceste stări pentru a:

- crește angajamentul (engagement),
- influența decizii rapide și impulsive,
- orienta utilizatorul spre acțiuni predefinite (vot, donație, protest, achiziție).

AI-ul nu creează emoția din nimic – ci o alimentează progresiv, oferind conținut care o întărește și o legitimează: postări alarmiste, știri negative, videoclipuri agresive sau mesaje care stârnesc indignarea morală.

Tehnici AI implicate:

- Emotion AI / machine learning afectiv – detectarea stării emoționale în timp real;
- Feed adaptiv emoțional – livrarea de conținut personalizat pentru a amplifica o emoție dominantă;
- Modelare comportamentală predictivă – identificarea momentelor când utilizatorul este mai vulnerabil (ex: târziu, după eșecuri, în criză personală);

[3] The Spectator - The real story of Cambridge Analytica and Brexit
https://www.spectator.co.uk/article/were-there-any-links-between-cambridge-analytica-russia-and-brexit/
The Gurdian - Cambridge Analytica did work for Leave.EU, emails confirm
https://www.theguardian.com/uk-news/2019/jul/30/cambridge-analytica-did-work-for-leave-eu-emails-confirm

- Recomandare algoritmică bazată pe stimulare emoțională (ex: momeală bazată pe furie, apel de frică, declanșează rușine).

Risc / Impact:

- Manipularea deciziilor sub presiunea emoțională – cumpărături compulsive, reacții violente, susținere necritică a unor cauze;
- Instabilitate psihologică – consum continuu de conținut negativ conduce la anxietate, depresie, paranoia informațională;
- Polarizare și ură colectivă – exploatarea furiei conduce la radicalizarea grupurilor și erodarea coeziunii sociale;
- Creșterea vulnerabilității în fața escrocheriilor și manipulărilor ideologice – emoțiile scad vigilența cognitivă.

Context de apariție:

- Campanii politice, crize sanitare, sociale, climatice;
- Scandaluri publice, tragedii, atacuri sau dezastre;
- Reclame agresive de tip „fear-based marketing”;
- Campanii de instigare sau „rage farming” pe rețele sociale.

Exemplu real:

În timpul pandemiei, milioane de utilizatori au fost expuși la conținut alarmist livrat de AI (ex: „vaccinul te va omorî”, „toți ne ascund adevărul”), pe baza istoricului lor de interacțiune. Sistemele au detectat că frica conduce la vizionări mai lungi, click-uri mai multe și (re)distribuiri în masă. Rezultatul: panică, neîncredere în autorități, dezbinare socială, acest aspect fiind surprins și în presă[4].

Indicatori de avertizare pentru utilizator:

- Simțim constant emoții negative intense după consumul digital (ex: furie, teamă, rușine, revoltă);
- Avem tendința să reacționăm imediat, fără să mai analizăm logic informația;
- Feed-ul nostru pare dominat de conținut negativ, catastrofic sau indignat;
- Observăm că toate mesajele „confirmă” o stare deja existentă de anxietate sau neîncredere.

Măsuri de prevenire / protecție:

- Limităm expunerea la conținut emoțional în perioade de vulnerabilitate personală;
- Ne antrenăm pentru a recunoaște „capcanele emoționale” algoritmice: titluri șocante, videoclipuri înșelător de dramatice, postări bazate pe urgență;
- Ne oferim timp între emoție și reacție – nu distribuim, comentăm sau acționăm imediat;
- Verificăm faptele din surse independente, mai ales când reacția noastră este una puternic emoțională;
- Ne antrenăm gândirea critică să detecteze „triggers” intenționate – dacă suntem prea furioși, probabil cineva și-a atins scopul.

[4] The Guardian - ‘Alarming’: convincing AI vaccine and vaping disinformation generated by Australian researchers
https://www.theguardian.com/australia-news/2023/nov/14/alarming-convincing-ai-vaccine-and-vaping-disinformation-generated-by-australian-researchers
The Trust & Safety Fundation - AI-Generated Disinformation Campaigns Surrounding COVID-19 in the DRC
https://www.trustandsafetyfoundation.org/blog/blog/ai-generated-disinformation-campaigns-surrounding-covid-19-in-the-drc

F. Atacuri de tip spear phishing automatizat cu conținut AI

„Nu mai este nevoie de un hacker priceput – AI-ul poate construi atacuri personalizate în masă, cu precizie letală.”

Descriere:

În acest scenariu, atacatorii folosesc inteligența artificială pentru a automatiza campanii de spear phishing – adică tentative de înșelăciune personalizate, direcționate către o anumită persoană sau un grup restrâns de potențiale victime. Spre deosebire de phishingul tradițional, care trimite mesaje generice, spear phishing-ul creat de AI este:

- Specific,
- Personalizat,
- Convingător,
- Adaptat contextului real al victimei.

AI-ul analizează prezența online a țintei (rețele sociale, articole, CV-uri, interacțiuni publice), generează mesaje personalizate perfect redactate și poate chiar simula conversații în timp real pentru a obține acces, date sau bani.

Tehnici AI implicate:

- Large Language Models (LLMs) – generarea de emailuri sau mesaje conversaționale adaptate la context (ex: ChatGPT, Claude, etc.);
- Profilare OSINT – AI-ul analizează date publice despre victimă (ex: loc de muncă, colegi, obiceiuri, interese);
- Simulare de identitate – imitarea stilului de scris al unui coleg, partener, client etc.;
- Voice cloning / deepfake audio – în unele cazuri, vocea unui superior este clonată pentru a comunica ordine false prin telefon.

Risc / Impact:

- Furt de identitate / date de autentificare – victimele oferă user, parolă, OTP sau semnează documente fără să știe;
- Compromiterea rețelelor interne ale unei organizații – acces prin social engineering către infrastructura IT;
- Fraudă financiară – ordine false de transfer bancar, plăți către conturi frauduloase;
- Discreditare personală sau profesională – folosirea conținutului obținut pentru șantaj, presiune sau distrugerea reputației.

Context de apariție:

- Companii (ex: angajați cheie, HR, contabilitate, IT);
- Jurnaliști, activiști, lideri politici;
- Persoane cu roluri administrative sau acces privilegiat în instituții;
- Campanii geopolitice, concurență comercială, atacuri APT.

Exemplu real:

În 2023, cercetători în securitate au demonstrat că un AI putea genera în sub 60 de secunde un email de spear phishing personalizat, aparent trimis de CEO-ul unei companii, folosind stilul său real de comunicare și făcând referire la proiecte interne reale (obținute din surse publice).

În teste, rata de accesare a fost de peste 70%, articolele din presă fiind destul de explicite în acest sens[5].

Indicatori de avertizare pentru utilizator:

- Primim un mesaj neobișnuit, dar bine scris, de la cineva cunoscut – cu un link, fișier sau solicitare urgentă;
- Contextul pare „la fix" – un proiect recent, o formulare familiară, un apel la autoritate;
- Se solicită acțiune rapidă, fără timp de verificare („urgent", „doar azi", „execută imediat");
- Orice ezitare este întâmpinată cu presiune pseudo-empatică: „înțeleg că ești ocupat, dar te rog...".

Măsuri de prevenire / protecție:

- Activăm autentificarea multifactor (MFA) pentru toate conturile importante;
- Verificăm întotdeauna solicitările suspecte printr-un canal alternativ (ex: apel direct, mesaj intern);
- Nu accesăm linkuri sau atașamente fără să verificăm adresa completă de expeditor și contextul solicitării;
- Folosim filtre anti-phishing, extensii de detecție AI și sisteme de protecție endpoint;
- Suntem conștienți că un mesaj foarte bine scris nu mai este o garanție a autenticității – ci poate reprezenta chiar semnalul unui atac într-un stadiu avansat.

G. Simulare de consens public prin rețele de conturi și boți AI

„Când mii de voci care par reale spun același lucru, ajungi să crezi că tu ești cel care greșește."

Descriere:

În acest scenariu, atacatorii sau actorii manipulatori folosesc rețele de conturi automatizate controlate de AI (boți sociali) pentru a crea iluzia unui consens social larg. Aceste conturi simulează persoane reale, cu profiluri credibile, imagini generate de AI, istorii de activitate și postări convingătoare.

Obiectivul este de a amplifica artificial o idee, o cauză, o indignare sau o direcție ideologică, astfel încât publicul larg să perceapă acel punct de vedere ca fiind:

- Majoritar,
- Rezonabil,
- Inevitabil.

Această presiune socială artificială înregistrează efecte semnificative asupra psihologiei individuale: dacă „toată lumea" pare să susțină ceva, e mai greu să rămâi în opoziție – sau chiar să mai pui întrebări.

Tehnici AI implicate:

- Generare de identități false (GANs) – imagini de profil ultra-realiste, dar complet artificiale;
- LLMs – generarea de postări, comentarii, reacții și mesaje aparent umane;

[5] Since Direct – Spear phishing attack
https://www.sciencedirect.com/topics/computer-science/spear-phishing-attack
MalwareBytes - AI-supported spear phishing fools more than 50% of targets
https://www.malwarebytes.com/blog/news/2025/01/ai-supported-spear-phishing-fools-more-than-50-of-targets

- Automatizare și orchestrare prin AI – controlul comportamentului simultan al mii sau chiar milioane de conturi (postare, redistribuire, atac, susținere);
- Manipulare conversațională – replici adaptate emoțional și logic, care simulează dezbateri între „oameni diferiți".

Risc / Impact:

- Fabricarea artificială a încrederii publice într-un produs, mesaj politic, teorie conspiraționistă sau atac la adresa unui grup;
- Marginalizarea vocilor reale – prin volum, repetitivitate și agresivitate, vocile opuse sunt acoperite sau descurajate;
- Constrângere socială indirectă – cei care nu aderă la „curentul majoritar" se autocenzurează sau își schimbă / ajustează opinia;
- Diluarea adevărului și a dezbaterii autentice – conversațiile online devin spații manipulate, fără pluralitate reală.

Context de apariție:

- Alegeri, referendumuri, crize politice;
- Campanii de propagandă internă sau externă;
- Promovarea de teorii conspiraționiste, pseudoștiință sau „produse minune";
- Dezinformare anti-occidentală, anti-UE, anti-NATO, anti-democratică (în context geopolitic).

Exemplu real:

În 2020, rețele sociale din mai multe țări au identificat mii de conturi coordonate, care promovau mesaje anti-vaccinare și anti-lockdown, pretinzând a fi părinți, medici, veterani sau cetățeni îngrijorați. Profilurile aveau imagini generate de AI și istorii false de activitate. Mesajul repetat: „poporul s-a trezit", mai multe analize fiind destul de clare în acest sens[6].

Indicatori de avertizare pentru utilizator:

- Multe comentarii identice sau foarte similare, postate în același timp;
- Profiluri recente, fără activitate autentică sau cu interacțiuni închise;
- Conturi care postează doar pe o temă unică, obsesiv, fără variații;
- Răspunsuri rapide, coordonate, ce „atacă" orice opinie diferită;
- Tendința ca „toată lumea să fie de acord" – fără nuanțe, critici sau dezbateri reale.

Măsuri de prevenire / protecție:

- Verificăm profilurile suspecte (ex: poze inversate, activitate, urmăritori, limbaj robotic);
- Nu ne lăsăm influențați de volum – întreabăm: „Cine sunt acești oameni? Există în realitate?";
- Suntem atenț la manipularea emoțională în masă – când „toți sunt indignați" ne întreabăm: de ce tocmai acum?
- Nu ne schimbăm convingerile doar pentru că „așa pare să creadă lumea" – cautăm argumente reale, nu doar aparențe.

[6] Comisia Europeană – Combaterea dezinformării
https://commission.europa.eu/strategy-and-policy/coronavirus-response/fighting-disinformation_ro
Centrul Euro-Atlantic pentru Reziliență - Barometrul rezilienței societale la dezinformare
https://e-arc.ro/wp-content/uploads/2022/05/Barometrul-rezilientei-societale-2022.pdf

H. Crearea de personaje publice artificiale pentru manipulare și influență

„Cine te influențează? Un om – sau o entitate creată în întregime de AI, cu o agendă precisă?”

Descriere:

În acest scenariu, inteligența artificială este exploatată pentru a construi de la zero personalități publice false (ex: influenceri, analiști, activiști, experți, „voci credibile”), controlate de operatori. Aceste personaje sunt dezvoltate cu:

- aspect vizual generat de AI (ex: fotografii realiste, avataruri animate),
- biografie convingătoare,
- conținut bine redactat (ex: texte, videoclipuri, postări),
- interacțiuni automate cu publicul.

Scopul este de a influența publicul, de a clădi încredere și de a injecta ulterior mesaje ideologice, comerciale sau manipulative în spațiul digital, fără riscul de a fi confruntat cu persoane reale sau riscul de pierdere a credibilității (pentru că nu există nimic de pierdut).

Tehnici AI implicate:

- Generare de imagini realiste (ex: GANs, StyleGAN, Midjourney) – pentru portrete, fotografii „de viață” etc.;
- LLMs (ex: ChatGPT, Claude, Mistral) – pentru postări, articole, comentarii, răspunsuri conversaționale;
- Voice synthesis și avataruri video (ex: Synthesia, D-ID) – pentru apariții „video” realiste și convingătoare;
- Social bot orchestration – conturi secundare automate care amplifică mesajele „personajului”.

Risc / Impact:

- Manipulare strategică de opinie – personajul câștigă încrederea și începe să distorsioneze adevărul referitor la subiecte sensibile;
- Crearea unor „lideri de opinie” fără responsabilitate sau fără identitate reală;
- Imitarea unor profesii cu autoritate (ex: doctori, avocați, jurnaliști, activiști umanitari);
- Influență geopolitică ascunsă – personaje aparent neutre ce promovează agende ostile;
- Interferență în spațiul civic, educațional, religios, medical sau politic.

Context de apariție:

- Platforme de social media, grupuri private, canale de video / streaming;
- Campanii de dezinformare sau rebranding ideologic;
- Promovarea de produse controversate, pseudoștiință, conspirații sau mișcări politice;
- Construirea de rețele „influente” controlate 100% de AI și algoritmi.

Exemplu real:

În 2022, o rețea de influenceri „femei de carieră” de pe Instagram promovau valori occidentale în regiunile din Orientul Mijlociu. Ulterior, s-a descoperit că toate erau avataruri AI controlate de o agenție guvernamentală, iar interacțiunile lor erau 100% generate automat – de la texte la răspunsuri directe și comentarii, mai multe articole fiind elaborate pe acest subiect[7].

[7] PC Tablet - Îmbrățișați valul digital: creșterea influențelor AI
https://pc-tablet.com/embrace-the-digital-wave-the-rise-of-ai-influencers/
You Dream AI - 10 exemple de influențatori AI pe Instagram (viitorul este aici)
https://yourdreamai.com/ai-influencer-examples-on-instagram/

Indicatori de avertizare pentru utilizator:

- Profiluri fără urme de activitate în afara platformei respective;
- Poze ”prea perfecte”, identice ca stil, cu fundaluri vagi sau imposibil de verificat;
- Informații biografice inconsistente sau imposibil de verificat;
- Ton ”excesiv de neutru”, constant și „corect”, fără fluctuații emoționale naturale;
- Activitate prea intensă (postări zilnice, reacții rapide, comentarii 24/7).

Măsuri de prevenire / protecție:

- Verificăm existența persoanei din alte surse (ex: cautăm în presă, evenimente, declarații publice autentice);
- Suntem suficient de sceptici când un „influencer nou” dispune rapid de un public numeros și un mesaj puternic repetitiv;
- Nu consideram „credibilitatea socială” (ex: like-uri, comentarii) ca dovadă de autenticitate;
- Rămânem atenți la puncelel în care „influencerul” începe să susțină idei partizane, extreme sau toxice – chiar dacă inițial părea echilibrat.

I. Campanii orchestrate prin aplicații mobile cu AI integrat (ex: fake news, mobilizare, radicalizare)

„Aplicația pare inofensivă. Dar în spate, AI-ul orchestrează o agendă invizibilă.”

Descriere:

În acest scenariu, o aplicație mobilă aparent benignă (ex: de știri, divertisment, educație, spiritualitate, comunitate, sau chiar de sănătate) integrează în fundal mecanisme AI de manipulare informațională. Aplicația devine o platformă de natură malignă, prin care sunt livrate conținuturi distorsionate, false sau ideologizate, cu scopul de a:

- Influența opinii,
- Dirija emoții,
- Mobiliza utilizatorii în acțiuni colective sau
- Radicaliza treptat anumite grupuri.

Pentru că aplicația este „de încredere” (ex: descărcată dintr-un magazin online oficial, bine evaluat, poate chiar sponsorizată de entități obscure dar legitime aparent), utilizatorul nu suspectează nimic.

Tehnici AI implicate:

- LLMs – generarea automată a conținutului în funcție de profilul și comportamentul utilizatorului;
- Feed personalizat controlat algoritmic – conținut adaptat în timp real pe baza reacțiilor;
- Emotion AI – detectarea dispoziției utilizatorului și adaptarea mesajelor în funcție de starea emoțională;
- Gamificare cu mecanici manipulative – premii, puncte sau mesaje emoționale care recompensează comportamente radicale sau obediente.

Risc / Impact:

- Răspândirea dezinformării sub forma „conținutului de încredere” – utilizatorul nu verifică sursa, deoarece percepe aplicația ca fiind sigură;
- Mobilizare pe teme false sau inflamatoare – proteste, acțiuni de masă, reacții colective;

- Radicalizare ideologică graduală – de la conținut „soft” la convingeri extreme, prin expunere repetitivă și adaptivă;
- Colectare masivă de date pentru profilare – comportamentale, sociale, politice sau religioase, fără consimțământ explicit;
- Construirea de comunități închise, autoreferențiale, greu de penetrat de mesajele alternative.

Context de apariție:

- Aplicații promovate ca fiind „alternative” la media tradițională („adevărul pe care alții îl ascund”);
- Aplicații pentru părinți, educație alternativă, spiritualitate, sănătate naturală;
- Platforme de chat sau rețele sociale noi, cu promisiuni de genul „libertate de exprimare totală”;

Campanii sponsorizate indirect de actori politici sau grupuri de influență netransparente.

Exemplu real:

În mai multe țări, aplicații mobile care pretindeau că oferă „știri necenzurate” au fost descoperite ca fiind controlate de rețele de propagandă partizane. Utilizatorii erau treptat expuși la narative false despre elite globale, conspirații sanitare sau apeluri la revoltă împotriva autorității. Totul, printr-o ”interfață curată” și cu o mască de profesionalism, așa cum a fost prezentată situația și în presă[8].

Indicatori de avertizare pentru utilizator:

- Aplicația oferă „adevăruri exclusive” sau promite să „te trezească”;
- Lipsa surselor clare sau invocarea constantă a unor „sisteme ascunse care ne mint”;
- Creșterea emoțiilor negative după utilizare (ex: frustrare, teamă, neîncredere totală în instituții);
- Recomandări care conduc spre grupuri închise, forumuri radicalizate sau apeluri la acțiune „urgentă”;
- Notificări frecvente, insistente, legate de crize, trădări, comploturi etc.

Măsuri de prevenire / protecție:

- Verificăm dezvoltatorul și sursa aplicației – cine o controlează, ce scopuri are;
- Verificăm dacă informațiile oferite sunt susținute și de alte surse independente;
- Suntem atenți la sentimentele generate de aplicație – dacă emoțiile negative sunt frecvente, acesta poate fi un semnal de manipulare;
- Dezinstalăm aplicațiile care ne oferă doar o singură perspectivă / doar un unghi de vedere și ne cultivă neîncrederea absolută față de orice altceva;
- Învățăm să identificăm narativele de control bazate pe frică, ură sau superioritate morală unilaterală.

J. Influențarea educației prin AI – resurse, platforme sau „mentori” care distorsionează adevărul

„Nu toate lecțiile vin din manuale – și nu toți profesorii sunt umani sau imparțiali.”

[8] Zimperium - Fake BBC News App: Analysis, https://zimperium.com/blog/fake-bbc-news-app-analysis

Descriere:

În acest scenariu, AI este folosit pentru a influența negativ formarea tinerilor, a elevilor sau a publicului general prin resurse educaționale distorsionate, părtinitoare sau complet false, livrate sub forma unor:

- Platforme de învățare,
- Aplicații educaționale,
- Chatbot-uri cu rol de „mentori" digitali,
- Videoclipuri didactice AI-generated,
- „Cursuri" alternative promovate ca fiind mai „autentice" decât cele din sistemul formal.

Aceste instrumente pot părea utile și inovatoare, dar sunt programate să includă intenționat narative manipulatoare, teorii nevalidate sau viziuni ideologice mascate drept „adevăruri ascunse".

Tehnici AI implicate:

- LLMs – generare automată de răspunsuri, explicații și lecții în funcție de întrebările elevului;
- Text-to-video + avatar AI – lecții video prezentate de „profesori" artificiali, convingători, dar inexistenți;
- Sisteme adaptive – care ajustează conținutul în funcție de nivelul, stilul cognitiv și emoțiile elevului;
- Microtargeting educațional – livrarea de „resurse" diferite pentru utilizatori cu profiluri ideologice diferite.

Risc / Impact:

- Modelarea incorectă a gândirii tinerilor – prin expunere repetată la informație manipulată, falsă sau ideologizată;
- Distrugerea încrederii în educația formală – în favoarea unor „sisteme alternative" controlate netransparent;
- Răspândirea conspirațiilor și pseudoștiinței sub pretext educațional;
- Polarizare ideologică timpurie – elevii ajung să fie programați să respingă automat anumite valori, teorii sau autorități științifice;
- Crearea de generații „antrenate" pentru obedință digitală, nu pentru gândire critică.

Context de apariție:

- Platforme de e-learning neacreditate, dar care devin populare în rândul tinerilor;
- Canale video de „cultură generală" cu agendă ascunsă;
- Aplicații mobile de „dezvoltare personală" care deviază spre dogmă sau activism radical;
- Chatboți de „ajutor la teme" care oferă răspunsuri părtinitoare, greșite sau speculative.

Exemplu real:

În 2023, UNESCO a tras un semnal de alarmă privind pericolul pe care inteligența artificială îl reprezintă pentru memoria colectivă a Holocaustului. Organizația a documentat modul în care anumite modele de AI – inclusiv motoare de căutare, sisteme conversaționale și instrumente generative – oferă rezultate inexacte, revizioniste sau profund distorsionate atunci când

utilizatorii caută informații despre Holocaust, acest aspect fiind publicat și pe site-ul oficial al instituției[9].

Indicatori de avertizare pentru utilizatori:

- Aplicația sau platforma are o abordare „alternativă”, dar respinge complet sistemele academice consacrate;
- Lecțiile conțin frecvent expresii ca „ce nu vrea nimeni să știi”, „adevărul ascuns” sau „manualele mint”;
- Răspunsurile AI sunt livrate cu ton autoritar, fără menționarea surselor;
- Profesorul digital promovează constant o anumită viziune ideologică, antiștiințifică sau conspiraționistă;
- Feedbackul nu încurajează gândirea critică, ci aderarea la o linie narativă prestabilită.

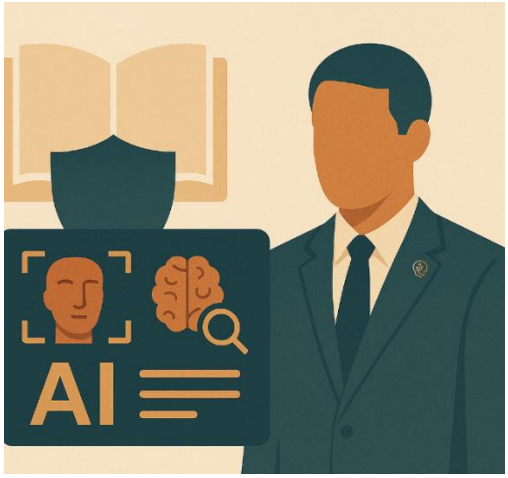
Măsuri de prevenire / protecție:

- Folosim platforme educaționale transparente, acreditate, cu conținut verificabil;
- Solicităm AI-ului surse pentru afirmațiile pe care le susține – și le verificăm;
- Nu folosim un singur canal educațional – comparăm răspunsurile și sursele între aplicații;
- Încurajeazăm întrebările, dezbaterea, îndoiala constructivă – nu acceptăm pasiv o „lecție predată”;
- Antrenăm elevii în educație media și AI literacy, pentru a recunoaște conținutul manipulator.

5 METODE DE PREVENIRE

Prevenția în era AI nu mai înseamnă doar instalarea unui antivirus sau evitarea site-urilor dubioase. Înseamnă cultivarea unor obiceiuri digitale sănătoase, dezvoltarea gândirii critice și înțelegerea modului în care funcționează algoritmii care ne însoțesc zilnic în mediul online. Nu ne confruntăm doar cu o problemă tehnologică, ci cu una cognitivă și socială.

În acest capitol, propunem o serie de măsuri practice – nu exhaustive, dar relevante – pentru utilizatori individuali, educatori, instituții și dezvoltatori de tehnologie.



[9] UNESCO - AI și Holocaustul: rescrierea istoriei? Impactul inteligenței artificiale asupra înțelegerii Holocaustului
https://www.unesco.org/en/articles/ai-and-holocaust-rewriting-history-impact-artificial-intelligence-understanding-holocaust

Scopul nu este doar de protecția pasivă, ci de dezvoltare a unei culturi privind vigilența digitală – în care fiecare utilizator devine activ, critic și conștient de mecanismele invizibile ce îi pot influența percepția și comportamentul.

5.1 Pentru utilizatorii individuali

„Nu ești neputincios în fața manipulării algoritmice – dar trebuie să înveți să recunoști și să reacționezi corect."

Inteligența artificială poate manipula subtil, convingător și invizibil. Dar publicul larg are la dispoziție metode concrete și eficiente pentru a-și proteja autonomia cognitivă și încrederea în realitate. Această secțiune prezintă un set de practici simple, dar importante, pentru a recunoaște, respinge și combate influența abuzivă a AI-ului asupra percepției și comportamentului.

A. Antrenează gândirea critică digitală

Primul și cel mai important filtru împotriva manipulării este propriul nostru discernământ.

- Întreabă-te mereu:
  - Cine vrea să cred asta?
  - Cine are de câștigat dacă reacționez emoțional sau impulsiv?
  - De ce apare această informație exact acum?
- Nu te baza pe prima impresie – AI-ul este antrenat să îți ofere exact acel conținut care „te prinde" rapid: titluri senzaționale, mesaje personalizate, imagini șocante. Învață să faci un pas înapoi și să îți regândești reacția.
- Antrenează-ți reflexul de analiză, nu doar de reacție. Gândirea critică este o formă de autoapărare digitală.

B. Verifică sursa și contextul conținutului

O informație fără sursă, fără context și fără autor verificabil e mai periculoasă decât o minciună asumată.

- Evită reacțiile emoționale spontane. Dacă ceva te face furios, anxios sau „pătruns de adevăr" instant, e un semnal că poate fi manipulativ.

- Verifică:
    - Cine a publicat conținutul?
    - Este autorul real, cunoscut, verificabil?
    - În ce context și când a apărut mesajul?
    - Cum a fost distribuit și cine l-a amplificat?
- Folosește surse externe, neutre, pentru confirmare. Nu te încrede doar în ce „îți apare" – ci caută activ răspunsuri alternative.

C. Recunoaște manipularea algoritmică

Dacă vezi mereu același tip de conținut, probabil nu ești informat – ești doar captiv într-un tipar digital.

- Dacă feed-ul tău pare „omogen" sau repetitiv, întreabă-te: *Unde sunt opiniile diferite? De ce nu le văd?*
- Caută activ contrariul:
    - Intră pe surse opuse ideologic,
    - Compară titlurile,
    - Discută cu oameni care au alte viziuni.
- Diversifică sursele:
    - Nu te informa dintr-o singură sursă;
    - Folosește motoare de căutare diferite, platforme independente, surse internaționale.
- Nu lăsa AI-ul să decidă ce vezi – preia tu controlul asupra propriei informări.

D. Folosește unelte de detectare AI / deepfake

Aparențele pot fi create de mașini. Fii mai vigilent decât un pixel.

- În fața unui videoclip, audio sau imagine dubioasă, folosește unelte specializate:
    - Deepware Scanner – detectează deepfake video/audio
    - Hive AI – analiză vizuală și auditivă automată
    - Sensity AI – soluții enterprise pentru detectarea manipulării vizuale
    - Microsoft Video Authenticator] – verifică autenticitatea videoclipurilor
- Caută inconsecvențe evidente:
    - Expresii faciale rigide sau artificiale,
    - Voci „plate" sau prea mecanice,
    - Sincronizare slabă a vorbirii,
    - Gesturi neverosimile, decoruri identice.
- Folosește funcția de reverse image search (ex: Google Images, Yandex) pentru a verifica, nu numai dacă o imagine a fost folosită anterior în alt context, ci și pentru manipulare și dezinformare

5.2 Pentru organizații

„Într-o eră în care reputația poate fi distrusă de un videoclip fals și deciziile interne pot fi influențate de un chatbot, organizațiile trebuie să se apere inteligent și proactiv."

Organizațiile (ex: companii, instituții publice, ONG-uri, structuri educaționale sau de securitate) sunt ținte prioritare în cadrul strategiilor de manipulare AI-based. Dezinformarea direcționată, fraudele conversaționale și compromiterea imaginii publice pot produce pierderi

importante financiare, operaționale, de încredere și de imagine. Prevenirea eficientă implică măsuri sistemice, tehnice și culturale.

A. Antrenamente de recunoaștere și prevenire a manipulării bazate pe AI

Instruirea echipelor reprezintă primul zid de apărare împotriva atacurilor cognitive și a tehnicilor sofisticate de inginerie socială.

- Programe interne de training pentru angajați, PR, HR, IT, juridic și top management privind:
  - Identificarea manipulării algoritmice;
  - Riscurile deepfake (ex: video, audio, text);
  - Recunoașterea phishingului conversațional și fraudelor cu mesaje AI.
- Simulări de atacuri cognitive:
  - Exerciții de testare a reacției la videoclipuri false cu „manageri”
  - Emailuri AI-scrise care simulează spear phishing,
  - Conversații simulate prin boți în scenarii de recrutare, solicitări financiare etc.
- Ghiduri interne de reacție rapidă:
  - Ce faci dacă apare un deepfake cu CEO-ul?
  - Cum validezi cereri urgente primite pe canale „credibile”?
  - Ce comunici public și cum gestionezi încrederea?

B. Politici de validare multisursă

Într-un context digital instabil, deciziile critice nu ar trebui validate doar printr-un singur canal (de comunicare).

- Orice decizie de natură financiară, contractuală sau strategică trebuie:
  - confirmată prin două sau mai multe canale independente (ex: e-mail + apel vocal + confirmare offline);
  - supusă unei autentificări încrucișate, mai ales dacă provine din surse neobișnuite sau în afara programului.
- Apelurile video și mesajele audio nu mai pot fi considerate dovezi solide, în contextul în care clonarea facială și vocală este accesibilă și realistă.
- Reactualizarea procedurilor interne astfel încât niciun departament să nu fie ”singur punct de decizie” în cazuri critice, fără o verificare multiplă.

C. Monitorizare reputațională automată și manuală

Reputația organizației este o țintă directă în războiul informațional. Un atac bine orchestrat poate distruge încrederea în câteva unități de timp (ore).

- Implementarea de soluții de monitorizare automată a mențiunilor brandului, numelor de lideri, produselor sau proiectelor, în special pe:
  - rețele sociale;
  - canale de mesagerie (ex: Telegram, WhatsApp groups);
  - surse alternative (ex: Dark Web, forumuri obscure);
  - platforme video și de fake-news.
- Detectarea campaniilor coordonate bazate pe AI:
  - postări simultane, conturi artificiale, replici identice;
  - deepfake-uri care simulează declarații ale reprezentanților companiei;

- documente fabricate ce „par” scurse din interior.
- Echipe de răspuns reputațional rapid: – contramăsuri proactive: identificarea și demontarea falsurilor;
  - campanii de informare transparente și directe pentru public, parteneri și presă.

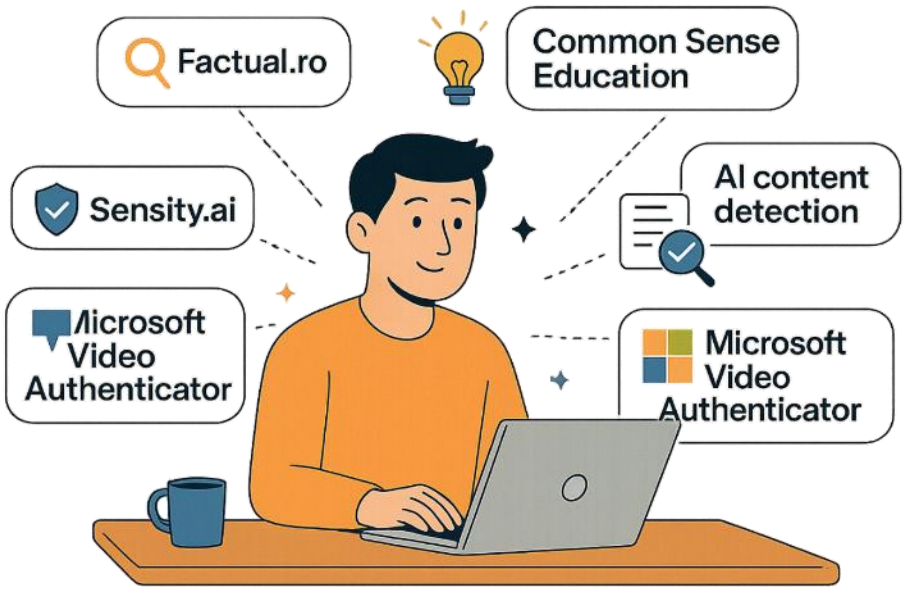
D. Colaborare cu experți, fact-checkeri și organizații specializate

Apărarea eficientă se construiește printr-un efort colectiv. Nimeni nu poate detecta, analiza și reacționa singur la valurile tot mai sofisticate de manipulare AI.

- Parteneriate active cu:
  - echipe de jurnaliști de investigație și verificare (fact-checking);
  - experți în securitate cibernetică, psihologie socială și comunicare de criză;
  - platforme specializate în detectarea AI (ex: Sensity, Deepware, Graphika);
  - ONG-uri care monitorizează spațiul informațional și propagarea falsurilor.
- Acces la rețele de alertare rapidă (inclusiv prin CERT-uri naționale sau rețele OSINT) pentru a răspunde în timp util în cazul campaniilor deepfake sau a atacurilor reputaționale.
- Participare la inițiative colective de educație și protecție împotriva manipulării digitale: cursuri, campanii publice, ghiduri de bune practici.

6 RESURSE ȘI ADRESE UTILE

„Accesul la informație corectă și la instrumente de verificare este prima linie de apărare împotriva manipulării bazate pe AI.”



Într-un peisaj informațional dominat tot mai mult de conținut generat de inteligența artificială, este important ca publicul larg, educatorii, profesioniștii și organizațiile să cunoască și să folosească resurse verificate și instrumente eficiente. Mai jos sunt disponibile câteva platforme utile în combaterea dezinformării, în educația critică și detectarea conținutului fals creat cu AI:

Verificări de informații pentru România - https://factual.ro

Platformă de fact-checking în limba română, dedicată combaterii declarațiilor false din spațiul public.

- Analizează și clasifică afirmațiile din politică, media și rețele sociale;
- Oferă surse, context și explicații pentru verdictul atribuit (ex: adevărat, fals, parțial adevărat etc.);
- Extrem de utilă pentru formarea reflexului de verificare a informației, mai ales în contexte electorale și sociale sensibile.

Analiză și detecție de conținut AI falsificat (deepfake, fake visual media) - https://sensity.ai

Platformă profesională de securitate vizuală, specializată în identificarea manipulărilor generate de AI.

- Detectează deepfake-uri video, imagini falsificate, voice cloning și fraude media vizuale;
- Oferă soluții avansate pentru organizații, media, instituții publice și corporații;
- Poate fi folositași în scop didactic, pentru a demonstra concret cum arată o manipulare vizuală.

Unelte AI experimentale oferite de Google - https://ai.google/tools

Colecție de aplicații și experimente bazate pe inteligență artificială, deschise publicului larg.

- Permite înțelegerea mecanismelor AI într-un mod interactiv și sigur;
- Include unelte pentru generare de texte, imagini, sunete și traducere automată;
- Utilă pentru cursuri introductive despre AI, alfabetizare digitală și analiză critică.

Campanii de manipulare analizate în Uniunea Europeană - https://www.euvsdisinfo.eu

Inițiativă a Serviciului European de Acțiune Externă (EEAS), dedicată expunerii și demontării campaniilor de dezinformare.

- Oferă o bază de date cu exemple de narative false, surse și canale de propagare;
- Analizează tematic și geografic modul în care dezinformarea afectează statele membre UE;
- Instrument important pentru jurnaliști, educatori, fact-checkeri și specialiști în comunicare strategică.

Resurse educaționale pentru gândire critică media - https://www.commonsense.org/education

Platformă non-profit cu resurse gratuite pentru educatori, părinți și elevi, axată pe dezvoltarea gândirii critice și a responsabilității digitale.

- Include lecții structurate despre fake news, bias media, influență socială și responsabilitate online;
- Adaptată pentru diferite vârste, cu materiale video, fișe de lucru și ghiduri pentru profesori;
- Poate fi integrată în activități curriculare sau extra curriculare dedicate educației media și AI.

Alte instrumente recomandate (pentru utilizare rapidă):

- InVID Plugin – extensie de browser pentru analiză video și imagistică;
- Deepware Scanner – verificare a autenticității fișierelor video / audio;
- NewsGuard – evaluare automată a credibilității site-urilor de știri;
- WhoTargetsMe – vizualizare și analiză a reclamelor politice direcționate pe social media.

Resurse educaționale recomandate

FBI - Federal Bureau of Investigation

AI Data Security – Best Practices

- https://media.defense.gov/2025/May/22/2003720601/-1/-1/0/CSI_AI_DATA_SECURITY.PDF

CISA Roadmap for Artificial Intelligence

- https://www.cisa.gov/sites/default/files/2025-04/ARCHIVE_20232024CISARoadmapAI508.pdf

AI Red Teaming: Applying Software TEVV for AI Evaluations

- https://www.cisa.gov/news-events/news/ai-red-teaming-applying-software-tevv-ai-evaluations

Serviciul Român de Informații – Centrul Național Cyberint

Intelligence

- https://intelligence.sri.ro/

Buletin Cyberint

- https://www.sri.ro/categorii/publicatii/

DNSC - Directoratul Național de Securitate Cibernetică

Deepfake și Inginerie Socială

- https://www.dnsc.ro/vezi/document/dnsc-ghid-inginerie-sociala
- https://www.dnsc.ro/vezi/document/dnsc-ghid-deepfake-organizatii

Detectare falsuri

- https://www.dnsc.ro/deepfake/

Poliția Română

Deepfake utilizat de infractorii cibernetici

- https://sigurantaonline.ro/deepfake-utilizat-de-infractorii-cibernetici-pentru-promovarea-unor-oportunitati-false-de-investitii-pe-retelele-sociale/

Test fraude online

- https://quiz.sigurantaonline.ro/

Asociația de Securitate Cibernetică pentru Cloud (CSA_RO – Chapter al CSA)

Responsabilități organizaționale în domeniul AI

- https://cloudsecurityalliance.org/artifacts/ai-organizational-responsibilities-ai-tools-and-applications

AI Controls Matrix

- https://cloudsecurityalliance.org/artifacts/ai-controls-matrix

Ghid strategic pentru implementarea AI

- https://cloudsecurityalliance.org/artifacts/dynamic-process-landscape-a-strategic-guide-to-successful-ai-implementation

Shadow Access and AI

- https://cloudsecurityalliance.org/artifacts/shadow-access-and-ai

Zero Trust and Artificial Intelligence Deployments

- https://cloudsecurityalliance.org/artifacts/confronting-shadow-access-risks-considerations-for-zero-trust-and-artificial-intelligence-deployments

Agentic AI Red Teaming Guide

- https://cloudsecurityalliance.org/artifacts/agentic-ai-red-teaming-guide

Clusterul de Excelență în Securitate Cibernetică

Cyber Security

- https://www.prodefence.ro/financial-fraud-fake-news-the-role-of-artificial-intelligence-in-disseminating-and-combating-false-information/

7 PREGĂTIRI PENTRU VIITORUL DEJA PREZENT

În acest nou ecosistem digital, riscul nu mai vine doar din dezinformare sau atacuri externe, ci și din expunerea constantă la conținut personalizat, emoțional și adesea manipulator. Tocmai de aceea, prevenția nu se referă doar la tehnologie, ci la o igienă digitală conștientă, cultivată zilnic.

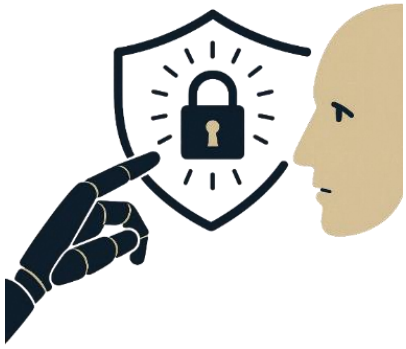
Pentru a face față acestei realități, sunt relevante următoarele cinci direcții fundamentale:

Educație digitală adaptată epocii AI

Educația clasică privind siguranța online trebuie să evolueze înspre o nouă paradigmă: alfabetizarea perceptivă digitală. Aceasta presupune formarea utilizatorilor – de la elevi până la decidenți – în recunoașterea manipulării subtile, identificarea conținutului generat de AI și înțelegerea modului în care algoritmii influențează atenția, emoțiile și convingerile.

În școli și instituții, vor fi necesare programe care includ:

- noțiuni despre personalizarea algoritmică;
- diferențierea între interacțiune umană și simulată;
- exerciții de analiză critică a surselor digitale



Reglementări clare și actualizate

Tehnologiile de generare automată a conținutului evoluează mult mai rapid decât legislația. Din acest motiv este importantă adoptarea unor cadre normative clare care să:

- interzică sau să reglementeze utilizarea conținutului manipulator creat de AI (ex: deepfake-uri în campanii electorale);
- oblige platformele să eticheteze transparent conținutul generat automat;
- impună responsabilitatea dezvoltatorilor AI pentru eventualele efecte negative ale aplicațiilor lor.

Aceste reglementări trebuie să protejeze atât drepturile individuale, cât și echilibrul democratic.

Transparență algoritmică și audit etic

Sistemele AI capabile să influențeze comportamentul uman (ex: platforme de social media, motoare de căutare, chatboți) trebuie să fie supuse unor audituri independente. Publicul are dreptul:

- să știe ce date personale sunt analizate;
- să înțeleagă de ce îi sunt afișate anumite tipuri de conținut;
- să opteze pentru un feed nemodificat algoritmic.

Totodată, instituțiile trebuie să susțină dezvoltarea unor standarde etice pentru AI, aplicate obligatoriu în educație, sănătate, justiție, politică etc.

Colaborare multidisciplinară

Problema manipulării perceptive nu este exclusiv tehnică. Avem nevoie de o abordare integrată în care să colaboreze:

- specialiști în securitate cibernetică și AI;
- psihologi și neurologi (pentru înțelegerea reacției emoționale);
- educatori și formatori (pentru diseminarea critică a informației);
- juriști și experți în drepturi digitale;
- eticieni și sociologi (pentru analiză de impact social).

Numai prin această colaborare vom putea înțelege efectele reale ale AI asupra societății și vom putea construi mecanisme eficiente de protecție.

Instrumente accesibile de detecție și verificare

La fel cum fiecare utilizator are acces la un motor de căutare sau un browser, în viitorul apropiat ar trebui să aibă acces și la:

- un instrument de detectare deepfake instalat pe telefon, laptop etc.;
- o extensie de browser care semnalizează conținutul generat de AI;
- o aplicație de verificare rapidă a sursei sau autenticității informației.

8 CONCLUZII

Inteligența artificială nu mai este o tehnologie în devenire, ci o forță invizibilă care structurează tot mai mult din ceea ce gândim, simțim și alegem. Într-un ecosistem digital hiperpersonalizat, unde conținutul este filtrat, emoțiile sunt măsurate, iar reacțiile sunt anticipate, riscul de influențare subtilă, dar sistematică, devine o realitate cu care ne confruntăm zilnic – adesea fără să știm.

Manipularea informațională nu mai arată ca în trecut. Nu este zgomotoasă, evidentă sau brută. Este fin calibrată, contextuală, personalizată – un algoritm care știe ce să spună, când să o spună și în ce ton, pentru a obține reacția dorită. Iar sursa acestor ajustări este, de multe ori, chiar comportamentul nostru digital: ce căutăm, ce ne reține atenția, ce ne sperie, ce ne consolează.

Această lucrare a urmărit să ofere o radiografie clară a modului în care AI-ul poate deveni un instrument de modelare perceptivă – prin tehnologie, prin psihologie, prin design conversațional. Mai mult decât un avertisment, ea propune și repere de igienă digitală și gândire critică, pentru a transforma utilizatorul pasiv într-un actor conștient al propriei realități informaționale.

9 GLOSAR DE TERMENI

Tehnologie și AI

- Inteligență artificială (IA / AI) – Simularea proceselor cognitive umane de către sisteme informatice capabile să învețe, să raționeze și să ia decizii autonome.
- Rețele neuronale artificiale (deep learning) – Arhitecturi de algoritmi inspirați de creierul uman, utilizate pentru recunoașterea de tipare complexe în texte, imagini sau voci.
- Machine Learning (învățare automată) – Ramură a AI care permite sistemelor să învețe și să se evolueze fără a fi programate explicit.
- Large Language Models (LLMs) – Modele de procesare a limbajului natural, antrenate pe volume mari de date pentru a genera și interpreta text (ex: ChatGPT, Gemini).
- Emotion AI / Machine learning afectiv – AI specializată în detectarea și interpretarea stărilor emoționale ale utilizatorilor.
- NLP (Natural Language Processing) – Tehnologii care permit înțelegerea și generarea limbajului uman de către mașini.
- GANs (Generative Adversarial Networks) – Rețele care pot genera imagini, sunete sau videoclipuri fals realiste.
- Face Swapping – Tehnică de înlocuire a feței unei persoane într-un video sau imagine cu fața altei persoane.
- Voice Cloning – Reproducerea artificială a vocii unei persoane reale cu ajutorul AI.
- Lip-syncing AI – Ajustarea mișcărilor buzelor într-un video pentru a corespunde unei voci generate sau schimbate.
- Synthmedia / Synth content – Conținut media creat integral de AI, fără intervenție umană.
- Avataruri sintetice / AI avatars – Reprezentări grafice animate generate de AI care pot imita persoane reale.
- Text-to-image models – AI care generează imagini pornind de la descrieri textuale.
- Motion capture AI – Tehnologii AI care reproduc mișcările corporale pentru animarea realistă a avatarurilor.
- Tacotron / WaveNet – Sisteme AI pentru sinteza vocii cu intonație și accent natural.
- Midjourney / DALL·E / Stable Diffusion / LLaMA / Claude / Gemini / ChatGPT / Mistral – Nume de modele AI avansate utilizate pentru generare de text, imagini sau conversații simulate.

Manipulare digitală

- Manipularea perceptivă – Influențarea invizibilă a modului în care o persoană percepe realitatea, prin conținut personalizat sau simulat.
- Manipulare algoritmică – Dirijarea comportamentului utilizatorului prin selecția automată a informațiilor afișate.
- Bula informațională – Spațiu digital personalizat în care utilizatorul primește doar informații care îi confirmă convingerile existente.
- Microtargeting psihografic – Livrarea de conținut emoțional personalizat, bazat pe profilul psihologic al utilizatorului.
- Spear phishing automatizat – Atacuri personalizate cu mesaje înșelătoare, generate de AI, care par a fi expedite de persoane cunoscute.
- Inginerie socială avansată – Folosirea AI pentru manipularea psihologică complexă în scopuri de fraudă, control sau influență.
- Recomandare comportamentală predictivă – Folosirea AI pentru a anticipa și modela deciziile utilizatorului.

- Filtrarea conținutului – Excluderea automată a punctelor de vedere alternative pentru a consolida o anumită percepție.
- Polarizare informațională – Separarea utilizatorilor în grupuri ideologice antagonice prin conținut direcționat.
- Radicalizare digitală – Proces prin care AI stimulează convingeri extreme prin expunere constantă la conținut radical.
- Simulare de consens social – Crearea artificială a impresiei că o opinie este susținută de majoritate.
- Simulare empatică / chatbot empatic – Chatboți care imită empatia umană pentru a obține încredere și influență.
- Influencer AI / influencer artificial – Conturi de rețea socială controlate de AI care simulează persoane reale pentru a genera influență.

Educație și securitate digitală

- Gândire critică digitală – Capacitatea de a analiza și evalua obiectiv conținutul digital.
- Educație cibernetică – Instruirea în privința riscurilor și mecanismelor de protecție online.
- Verificare factuală – Procesul de analiză a unei informații pentru a-i confirma veridicitatea.
- Audit algoritmic – Evaluarea sistematică a modului în care funcționează și influențează un algoritm.
- Transparență algoritmică – Dreptul utilizatorului de a ști cum sunt prelucrate datele și de ce primește un anumit conținut.
- Reflex de verificare – Reacția automatizată de a valida informația înainte de a o crede sau distribui.
- Igienă informațională – Ansamblu de practici pentru menținerea unui consum de informație sănătos și echilibrat.
- Autoapărare informațională – Set de cunoștințe și tehnici prin care utilizatorul se protejează de manipulare și dezinformare.
- Conținut generat de AI (AI-generated content) – Orice material creat automat de o inteligență artificială.
- Etică AI – Ramură care analizează implicațiile morale ale dezvoltării și utilizării inteligenței artificiale

10 BIBLIOGRAFIE

Associated Press. (2023). AI tools can fabricate disinformation easily. https://www.apnews.com/article/afb4618ff593db9e3e51ecbd91dc3eef

Bitdefender - Atenție la escrocherii la angajare, https://www.bitdefender.com/en-us/blog/hotforsecurity/8-telegram-scams-how-not-to-get-scammed

Centrul Euro-Atlantic pentru Reziliență - Barometrul rezilienței societale la dezinformare, https://e-arc.ro/wp-content/uploads/2022/05/Barometrul-rezilientei-societale-2022.pdf

EuroNews - Oferte de locuri de muncă false, https://www.euronews.com/next/2023/10/23/behind-the-global-scam-worth-an-estimated-100m-targeting-whatsapp-users-with-fake-job-offe

Europa Liberă România. (2022). România și cenzura internetului. https://romania.europalibera.org/a/romania-si-cenzura-internetului/32092813.html

European Commission. (2020). Fighting coronavirus disinformation. https://commission.europa.eu/strategy-and-policy/coronavirus-response/fighting-disinformation_ro/

EUvsDisinfo. (n.d.). Fighting disinformation. https://www.euvsdisinfo.eu

Federal Trade Commission. (2023, July). Job offer through Telegram Messenger? Not so fast. https://consumer.ftc.gov/consumer-alerts/2023/07/job-offer-through-telegram-messenger-not-so-fast

Financial Times. (2023). AI-generated spear phishing emails target executives. https://www.ft.com/content/d60fb4fb-cb85-4df7-b246-ec3d08260e6f/

France24 - Debunking a deepfake video of Zelensky telling Ukrainians to surrende, https://www.france24.com/en/tv-shows/truth-or-fake/20220317-deepfake-video-of-zelensky-telling-ukrainians-to-surrender-debunked

Graphika. (n.d.). Reports. https://graphika.com/reports

Hao, K. (2023, November 3). How fake news apps spread disinformation under the radar. MIT Technology Review. https://www.technologyreview.com/2023/11/03/apps-disinformation-misinformation-ai

Hart, K. (2021, February 23). Memes misinformation and coronavirus. Axios. https://axios.com/2021/02/23/memes-misinformation-coronavirus-56/

House of Commons Digital, Culture, Media and Sport Committee. (2019). Disinformation and 'fake news': Final Report. UK Parliament. https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

MalwareBytes - AI-supported spear phishing fools more than 50% of targets. https://www.malwarebytes.com/blog/news/2025/01/ai-supported-spear-phishing-fools-more-than-50-of-targets

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. Nature Human Behaviour, 1(9), 1-6. https://www.nature.com/articles/s41562-017-0099/

NewsGuard. (2023). AI-generated content tracker. https://www.newsguardtech.com/special-reports/ai-generated-content-tracker

PC Tablet - Îmbrățișați valul digital: creșterea influențelor AI, https://pc-tablet.com/embrace-the-digital-wave-the-rise-of-ai-influencers/

Persily, N. (2018). Digital Influence and Political Microtargeting. Journal of Democracy, 29(2), 64–78.
https://muse.jhu.edu/article/690796/

Prodefence – A. Anghelus (2024, August). Financial Fraud & Fake News: The Role of Artificial Intelligence in disseminating and combating false information. https://www.prodefence.ro/financial-fraud-fake-news-the-role-of-artificial-intelligence-in-disseminating-and-combating-false-information/

Reuters - Deepfake footage purports to show Ukrainian president capitulating, https://www.reuters.com/world/europe/deepfake-footage-purports-show-ukrainian-president-capitulating-2022-03-16/

Roozenbeek, J., van der Linden, S., & Nygren, T. (2022). Exposure to online misinformation about COVID-19 and vaccine hesitancy. Scientific Reports, 12, Article 10070. https://www.nature.com/articles/s41598-022-10070-w/

SecurityWeek. (2023). AI now outsmarts humans in spear phishing – analysis shows. https://www.securityweek.com/ai-now-outsmarts-humans-in-spear-phishing-analysis-shows/

Since Direct – Spear phishing attack, https://www.sciencedirect.com/topics/computer-science/spear-phishing-attack

SoSafe Awareness. (2023). One in five people click on AI-generated phishing emails https://sosafe-awareness.com/company/press/one-in-five-people-click-on-ai-generated-phishing-emails-sosafe-data-reveals

The Guardian - 'Alarming': convincing AI vaccine and vaping disinformation generated by Australian researchers, https://www.theguardian.com/australia-news/2023/nov/14/alarming-convincing-ai-vaccine-and-vaping-disinformation-generated-by-australian-researchers

The Gurdian - Cambridge Analytica did work for Leave.EU, emails confirm, https://www.theguardian.com/uk-news/2019/jul/30/cambridge-analytica-did-work-for-leave-eu-emails-confirm

The Spectator - The real story of Cambridge Analytica and Brexit, https://www.spectator.co.uk/article/were-there-any-links-between-cambridge-analytica-russia-and-brexit/

The Trust & Safety Fundation - AI-Generated Disinformation Campaigns Surrounding COVID-19 in the DRC, https://trustandsafetyfoundation.org/blog/ai-generated-disinformation-campaigns-surrounding-covid-19-in-the-drc/

Timberg, C., & Tiku, N. (2023, December 17). AI-generated fake news sites multiply online, spreading misinformation. The Washington Post. https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation

UNESCO - AI și Holocaustul: rescrierea istoriei? Impactul inteligenței artificiale asupra înțelegerii Holocaustului, https://www.unesco.org/en/articles/ai-and-holocaust-rewriting-history-impact-artificial-intelligence-understanding-holocaust

You Dream AI - 10 exemple de influențatori AI pe Instagram (viitorul este aici), https://yourdreamai.com/ai-influencer-examples-on-instagram/

Zimperium - Fake BBC News App: Analysis, https://zimperium.com/blog/fake-bbc-news-app-analysis

ANALIZEAZĂ – DECIDE – ACȚIONEAZĂ