

Inteligența Artificială & Manipularea Percepției Umane

Programul educațional
"Analizează – Decide - Acționează"

AUTORI

Mircea Constantin ȘCHEAU

Președinte – Asociația de Securitate Cibernetică pentru Cloud

Alexandru Ciprian ANGHELUȘ

Președinte – Clusterul de Excelență în Securitate

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 ciberneticice

Într-o eră digitală aflată într-o transformare accelerată, unde tehnologia aduce atât oportunități, cât și amenințări, securitatea utilizatorilor devine o prioritate. Atacurile ciberneticice 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.

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 fraudă, 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.

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



Cyber Security
Cluster of Excellence

CUPRINS

1.	NOȚIUNI GENERALE	7
1.1	Ce este Inteligența Artificială	7
1.2	Percepția umană și Inteligența Artificială.....	7
1.3	De ce este importantă înțelegerea mecanismelor din spatele AI-ului.....	8
2	ELEMENTE TEHNICE ȘI MECANISME	9
2.1	Tehnologii implicate	10
A.	Rețele neuronale artificiale (deep learning).....	10
B.	Large Language Models (LLMs) (ex: ChatGPT, Gemini, Claude, Mistral etc).....	13
C.	Machine Learning Afectiv (Emotion AI) – detectarea și manipularea emoțiilor	14
D.	AI vizual – imagini, video, deepfake, avataruri sintetice	16
2.2	Mecanisme de manipulare perceptivă.....	17
A.	Algoritmi de știri (feed) personalizat.....	17
B.	Microtargeting psihografic – influențarea personalizată a percepției și comportamentului.....	19
C.	Generare de conținut fals credibil – iluzia realității algoritmice.....	20
D.	Automatizarea conversației și manipulării prin boți avansați.....	21
3	MANIPULAREA PERCEPȚIEI CU AJUTORUL INTELIGENȚEI ARTIFICIALE....	22
3.1	Definirea contextului	23
3.2	Mecanisme de captare a atenției utilizatorului și pentru manipularea cognitivă a acestuia	24
A.	Filtrarea conținutului – ascunderea perspectivelor alternative	25
B.	Clasificarea emoțională a utilizatorului – generarea de conținut adaptat stării emoționale	26
C.	Crearea de conținut fals credibil – afectarea percepției realității.....	27
D.	Simularea empatiei și încrederii – obținerea ascultării și influențarea sentimentului de loialitate	28
E.	Recomandare comportamentală predictivă – modelarea deciziilor utilizatorului.....	30
4	AI ÎN INGINERIA SOCIALĂ ȘI DEZINFORMARE	31
4.1	Utilizarea în scopuri malițioase	32
4.2	Scenarii de utilizare.....	33
A.	Bula informațională asistată de Inteligența Artificială	33
B.	Boți conversaționali pentru fraudă sau recrutare falsă.....	34
C.	Generare de materiale false hiperrealiste (deepfake).....	35
D.	Mesaje personalizate de influență (microtargeting AI)	37
E.	Declanșarea controlată a emoțiilor (exploatarea emoțiilor negative de către AI).....	38
F.	Atacuri de tip spear phishing automatizat cu conținut AI.....	40

G. Simulare de consens public prin rețele de conturi și boți AI.....	41
H. Crearea de personaje publice artificiale pentru manipulare și influență.....	43
I. Campanii orchestrate prin aplicații mobile cu AI integrat (ex: fake news, mobilizare, radicalizare)	44
J. Influențarea educației prin AI – resurse, platforme sau „mentori” care distorsionează adevărul	45
5 METODE DE PREVENIRE	47
5.1 Pentru utilizatorii individuali	48
A. Antrenează gândirea critică digitală.....	48
B. Verifică sursa și contextul conținutului	48
C. Recunoaște manipularea algoritmică	49
D. Folosește unelte de detectare AI / deepfake.....	49
5.2 Pentru organizații	49
A. Antrenamente de recunoaștere și prevenire a manipulării bazate pe AI.....	50
B. Politici de validare multisursă.....	50
C. Monitorizare reputațională automată și manuală	50
D. Colaborare cu experți, fact-checkeri și organizații specializate.....	51
6 RESURSE ȘI ADRESE UTILE	51
7 PREGĂTIRI PENTRU VIITORUL DEJA PREZENT.....	54
8 CONCLUZII.....	55
9 GLOSAR DE TERMENI	56
10 BIBLIOGRAFIE	58

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 algoritmi 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 conștientizarea 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:

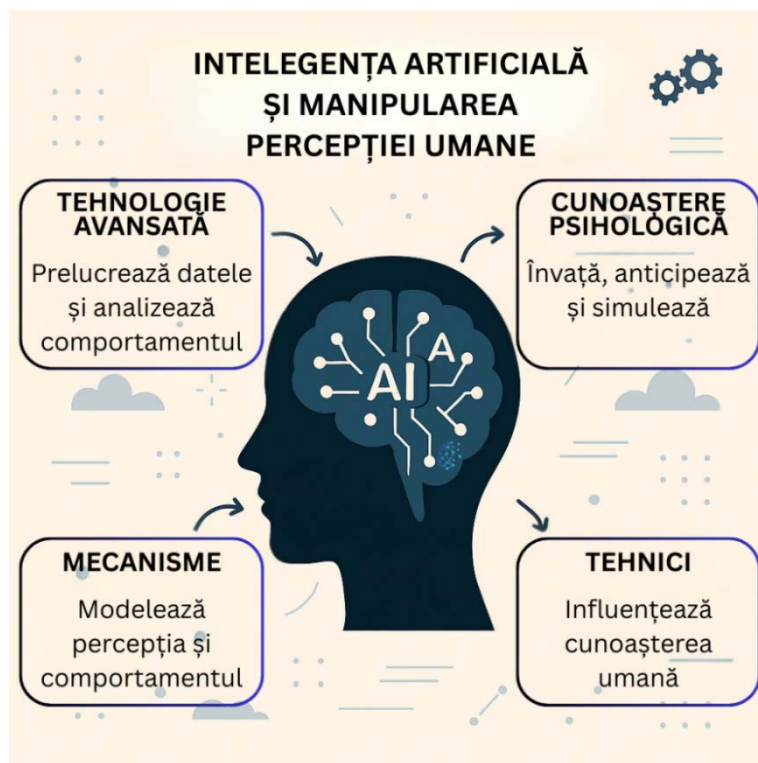
- Ce este un algoritm de recomandare și cum influențează ceea ce văd?
- Cum poate o rețea neuronală să înțeleagă emoțiile mele?
- Ce se întâmplă cu datele pe care le ofer, conștient sau inconștient?
- De ce anumite conținuturi par „făcute exact pentru mine”?

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 „învăță” 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 deturneză);
- 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 ”injecțai” 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, algoritmiile 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 distribuți, 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 țărțat(i) 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?

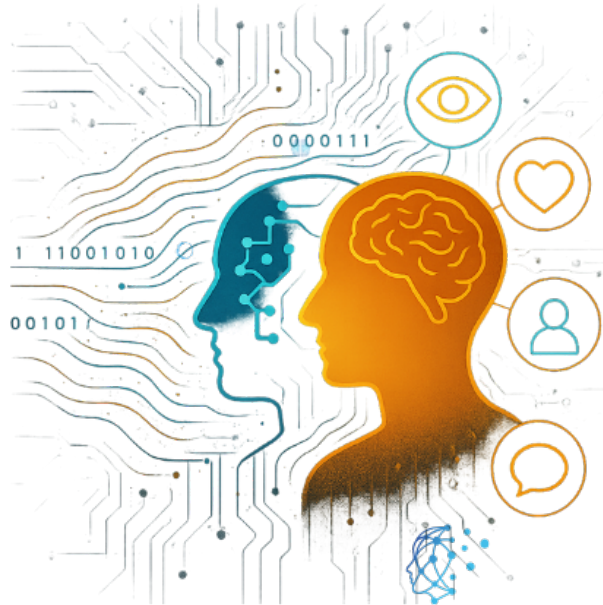
- Folosim scepticismul activ în conversațiile online – punem întrebări concrete, cerem dovezi ale identității, nu acționăm impulsiv;
- Verificăm întotdeauna sursa și scopul unei conversații, mai ales în cazul ofertelor, cererilor de ajutor sau promisiunilor de câștiguri rapide;
- Folosim site-uri verificate (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 algoritmi 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 - algoritmi 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 conștientizeze influența.

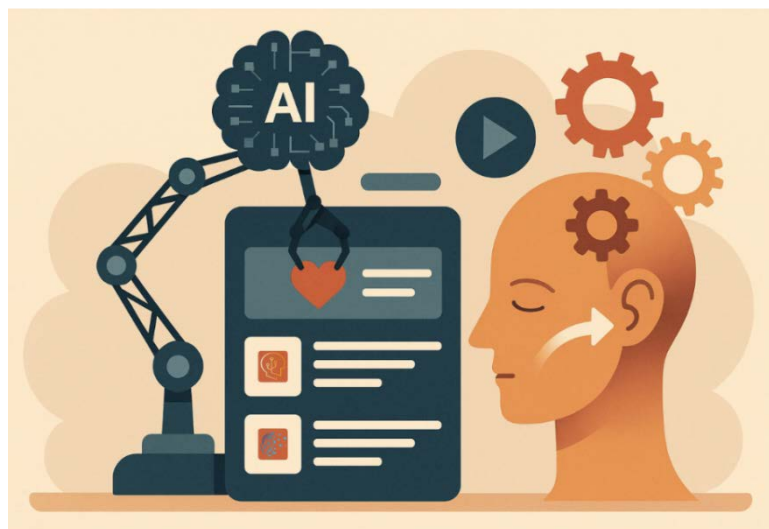
Exemple de manifestare

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 *affective computing* și este 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 interacțiilor 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) și 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, mișcărilor 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, intimidare 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 oră 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 „prietenosi” 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: „Vorbești 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?

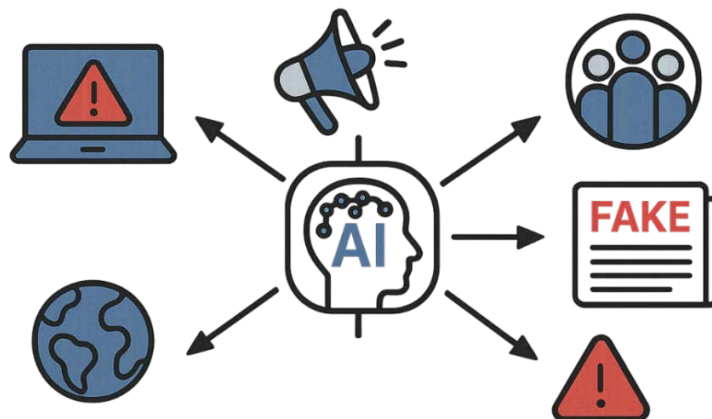
- Limităm partajarea de date comportamentale (dezactivăm istoricul, limităm cookie-uri, folosim browsere securizate).
- Conștientizăm că AI-ul cunoaște tiparele noastre mai bine decât noi înșine.
- Nu luăm decizii importante în momente de stres, oboseală sau impuls emoțional. Întrebăm: „Asta îmi doresc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

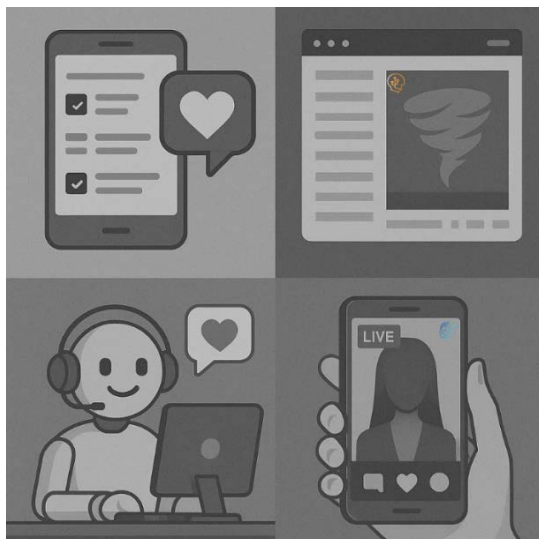
În trecut, ingineria socială se baza pe cunoștințe relativ nefundamentate referitoare la psihologie și pe intervenții generale. Astăzi, însă, AI permite atacuri scalabile, personalizate, automate și perfect sincronizate, fără contact fizic și adesea fără ca victimele să conștientizeze că au fost vizate.

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 aliază” 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 recunoaștem 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 algoritmiile 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 deja².

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 surrender
<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 vot³.

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.

Măsurile 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 Guardian - 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 omori”, „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)distribuire în masă. Rezultatul: panică, neîncredere în autorități, dezbinare socială, acest aspect fiind surprins și în presă⁴.

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/alarmed-convincing-ai-vaccine-and-vaping-disinformation-generated-by-australian-researchers>

The Trust & Safety Foundation - 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 sens⁵.

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...”.

Măsurile 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, rețelele 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 sens⁶.

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 – întrebăm: „Cine sunt acești oameni? Există în realitate?”;
- Suntem atenți la manipularea emoțională în masă – când „toți sunt indignați” ne întrebăm: 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 umani);
- 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 „influențe” 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 subiect⁷.

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).

Măsurile 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 considerăm „credibilitatea socială” (ex: like-uri, comentarii) ca dovadă de autenticitate;
- Rămânem atenți la punctele 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 opiniei,
- 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 presă⁸.

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.

Măsurile 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 manipulative, 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 obediență 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 instituției⁹.

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.

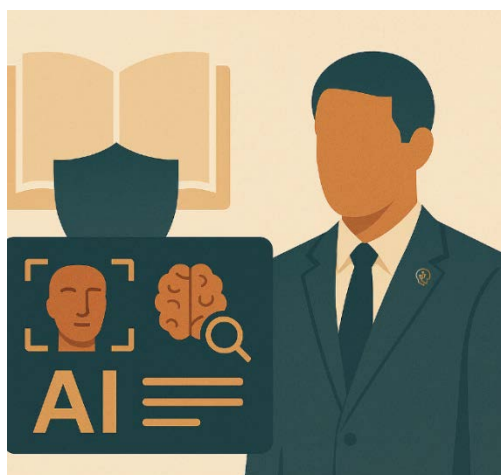
Măsuri de prevenire / protecție:

- 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;
- Încurajăm întrebările, dezbateră, 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 algoritmi 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-scrite 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 comunică 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.

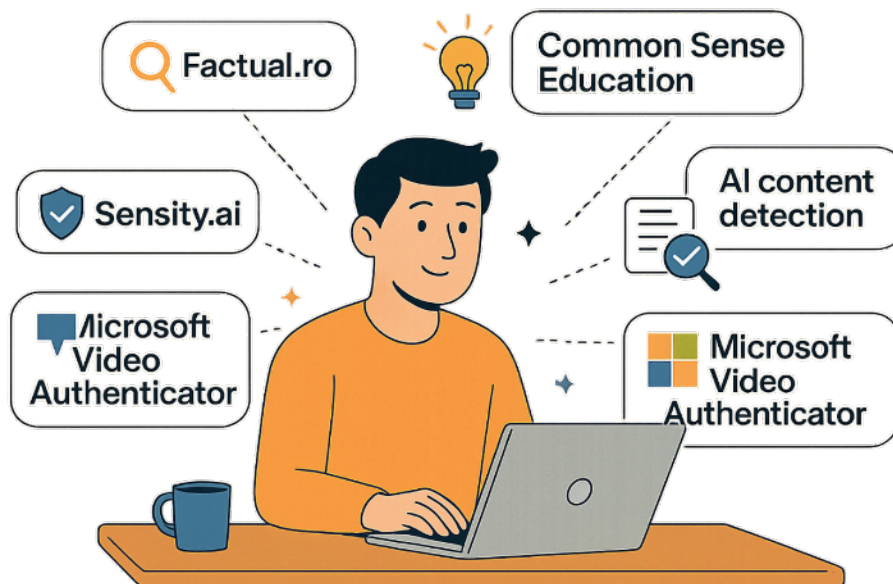
D. Colaborare cu experți, fact-checkeri și organizații specializate

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, voce clonaj și fraude media vizuale;
- Oferă soluții avansate pentru organizații, media, instituții publice și corporații;
- Poate fi folosită 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 narrative 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 algoritmi 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 surrender, <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/cmcmds/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 Foundation - 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>

Artificial Intelligence & The Manipulation of Human Perception

The educational program
"Analyze – Decide – Act"

AUTORS

Mircea Constantin ȘCHEAU

President – Cloud Security Alliance Romanian Chapter

Alexandru Ciprian ANGHELUȘ

President – Cyber Security Cluster of Excellence

CYBER EDUCATION THE FOUNDATION OF DIGITAL PROTECTION

In a world where technology evolves faster than users' ability to understand its risks, cyber education is becoming one of the key pillars of both personal and organizational safety. Whether we're talking about individual users, employees, students, or decision-makers, we must acknowledge that we are all exposed daily to increasingly sophisticated digital threats..

Knowledge is both weapon and armor.
Prevention begins with understanding.

Through continuous training, vigilance, and the practical application of knowledge, we can reduce the impact of attacks and protect our shared values: trust, privacy, and integrity.

ADA Educational Program – “Analyze – Decide – Act”

Critical Thinking, Digital Responsibility, and Active Protection in the Age of Cyber Risks

In a rapidly evolving digital era, where technology brings both opportunities and threats, user security becomes a top priority. Modern cyberattacks exploit not only technical vulnerabilities but especially human ones: lack of awareness, absence of vigilance, information overload, or misplaced trust in appearances.

“**Analyze – Decide – Act**” is an educational program aimed at strengthening both personal and collective digital resilience through awareness, hands-on training, and the development of critical thinking tailored to new forms of informational aggression.

- The program's series of materials targets:
- the general public (adults, seniors, parents, youth),
- children and teenagers (in educational contexts),
- public servants and professionals,
- decision-makers and leadership personnel.

The program's goal is to turn information into a tool of defense and transform the user from a passive target into an aware and active actor in the face of digital threats.

Through these materials, we aim to:

- increase individual and institutional awareness and caution;
- reduce the impact of digital attacks through education and rapid response;
- promote a culture of reporting, collaboration, and solidarity between users and specialists;
- foster responsible digital behavior aligned with current policies and regulations.

This project is a sustained effort in cyber literacy, built on principles of accessibility, applicability, and continuous updating, in order to respond to the dynamic nature of real-world risks in the virtual space.

PARTNERS



DIRECTORATUL NAȚIONAL
DE SECURITATE CIBERNETICĂ



Summary

This paper explores the ways in which Artificial Intelligence (AI) is exploited to influence perception and distort reality. The phenomenon is analyzed through the lens of algorithmic mechanisms such as personalization, the generation of credible false content (e.g., deepfakes, voice cloning, AI-generated text), conversational stimulation, and predictive emotional modeling.

The scenarios presented — involving disinformation techniques, ideological manipulation within artificial social constructs, emotional fraud, and automated conversational attacks — serve illustrative purposes. The referenced technologies (e.g., LLMs, GANs, Emotion AI, personalized feeds, microtargeting) and associated systemic risks are relevant as of the time of this material's development.

Beyond its descriptive component, the paper provides a necessary framework for developing critical thinking, presenting concrete methods for detection, prevention, and response. These are intended for the general public, institutions, and educational professionals.

Thus, the document serves as a tool for raising awareness about highly personalized, automated, and difficult-to-detect forms of social engineering.

Keywords

Artificial Intelligence, manipulation, digital disinformation, algorithms, deepfake, spear phishing, education, media, security.

Message to Readers

Artificial Intelligence is neither good nor bad.

It is a tool - a powerful one. Capable of learning from the information we provide, reacting to triggers and producing outcomes based on the goals of those who control it.

In the right hands, AI can save lives, enhance education, combat fraud, prevent cyberattacks, and support the development of society. In criminal — or merely irresponsible — hands, the same technology can be weaponized for manipulation, control, deception, ideological programming, and social destabilization.

This is why education is one of the most effective forms of protection.

If we understand how it works, we can more easily recognize when and how it is being used against us.

This guide is not intended to frighten, but to prepare — and preparation begins with a simple truth:

*Artificial Intelligence is a tool.
The power — and the danger — lie in the hands of those who wield it.*

** This document includes technical terms and standard designations, so that all readers may become familiar with this information.*



Cyber Security
Cluster of Excellence

TABLE OF CONTENTS

1.	GENERAL CONCEPTS	7
1.1	What is Artificial Intelligence.....	7
1.2	Human perception and Artificial Intelligence.....	7
1.3	Why understanding the mechanisms behind AI is important	8
2	TECHNICAL ELEMENTS AND MECHANISMS	9
2.1	Technology involved	10
	A. Artificial Neural Networks (Deep Learning).....	10
	B. Large Language Models (LLMs).....	13
	C. Affective Machine Learning (Emotion AI)	14
	D. Visual AI – images, video, deepfakes, synthetic avatars.....	166
2.2	Mechanisms of perceptual manipulation	17
	A. Personalized news (feed) algorithms	17
	B. Psychographic microtargeting – personalized influence on perception and behavior.....	18
	C. Credible fake content generation – the illusion of algorithmic reality	20
	D. Automated conversation and manipulation through advanced bots	21
3	MANIPULATING PERCEPTION THROUGH ARTIFICIAL INTELLIGENCE.....	22
3.1	Defining the context.....	23
3.2	Mechanisms for capturing user attention and cognitive manipulation	24
	A. Content filtering – hiding alternative perspectives	24
	B. Emotional classification – generating content based on the user's emotional state.....	25
	C. Credible fake content generation – distorting the perception of reality.....	26
	D. Simulated empathy and trust – gaining compliance and influencing loyalty	27
	E. Predictive behavioral recommendation – modeling user decisions	29
4	AI IN SOCIAL ENGINEERING AND DISINFORMATION	30
4.1	Malicious use cases.....	31
4.2	Use scenarios	31
	A. The AI-powered information bubble	32
	B. Conversational bots for fraud or fake recruitment	33
	C. Generation of hyper realistic fake media (Deepfakes)	34
	D. AI-powered personalized influence messaging (Microtargeting)	36
	E. Emotionally triggered manipulation (AI-driven exploitation of negative emotions) ..	37
	F. Automated AI-powered spear phishing attacks.....	39
	G. Simulated public consensus via AI bot networks	40
	H. Artificial public personalities for influence and manipulation	41
	I. Orchestrated campaigns via mobile apps with embedded AI	43

J. AI-Based influence in education – platforms, “mentors,” and distorted learning resources	44
5 PREVENTION METHODS.....	46
5.1 For individual users.....	46
A. Train your digital critical thinking.....	47
B. Check the source and context of the content	47
C. Recognize algorithmic anipulation	47
D. Use AI / deepfake detection tools	47
5.2 For organizations	48
A. AI manipulation awareness & defense training	48
B. Multi-channel validation policies	48
C. Automated and manual reputation monitoring	49
D. Collaborate with experts, fact-checkers & specialized organizations	49
6 USEFUL RESOURCES AND ADDRESSES	50
7 PREPARING FOR THE ALREADY-PRESENT FUTURE.....	52
8 CONCLUSIONS	54
9 GLOSSARY	55
10 BIBLIOGRAPHY	56

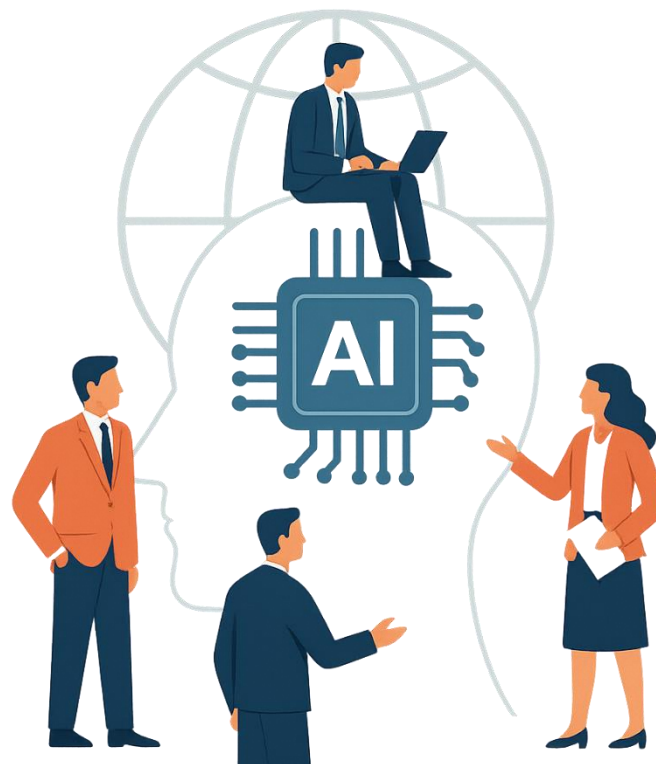
1. GENERAL CONCEPTS

1.1 What is Artificial Intelligence

Artificial Intelligence (AI) refers to a set of computer technologies capable of simulating human thought processes — such as learning, reasoning, perception, and decision-making. The most well-known types include machine learning, deep neural networks (deep learning), natural language processing (NLP), visual recognition, and generative models (e.g., ChatGPT, DALL·E, Gemini, Claude, etc.).

Unlike traditional algorithms, modern AI does not follow a fixed set of rules but instead “learns” from data sets and adapts to human behavior. Advanced models — especially generative ones — can create text, images, voices, or even videos that are nearly indistinguishable from real content.

These capabilities bring extraordinary benefits — from automation to education and research — but also carry significant risks, particularly in the areas of information manipulation, trust, and human emotions.



1.2 Human perception and Artificial Intelligence

Artificial Intelligence (AI) is no longer just a tool of the future — it is an active part of our digital present. Whether we're watching a video on a streaming platform, reading news online, or interacting with a virtual assistant, there is a high probability that an AI algorithm is running in the background, deciding what we see, what we hear, and even how we interpret reality.

At the core of these processes lies AI's ability to analyze human behavior, learn from collected data, and generate content or responses that imitate or stimulate authentic human reactions. These capabilities have valuable applications in fields such as medicine, education, and

automation. However, there is a dangerous flipside: the risk of large-scale informational and emotional manipulation.

Unlike traditional influence methods (e.g., advertising, propaganda, social persuasion), AI introduces a new level of precision and subtlety in manipulation. Modern algorithms can identify emotional vulnerabilities, behavioral patterns, and users' psychological preferences with astonishing accuracy, generating personalized content that triggers strong emotions and impulsive decisions — often without the target being aware of the process.

From content recommendations that reinforce existing beliefs and isolate users in information bubbles, to the highly realistic simulation of real people (through deepfakes, voice cloning, or AI avatars), artificial intelligence becomes an active vector in shaping perception — and by extension, the subjective reality of everyone.

This emerging reality raises critical questions:

- How can we recognize content that has been generated or manipulated by AI?
- Where is the line between helpful recommendation and intentional manipulation?
- What does truth mean in an age where any voice, image, or activity can be artificially replicated with precision?

To address these challenges, a comprehensive cyber education is needed — one that goes beyond basic concepts and addresses the cognitive, psychological, and social dimensions of interacting with intelligent technologies.

This educational guide aims to:

- Explain how AI can influence human perception, through both visible and subtle mechanisms
- Analyze risks and technologies, offering a realistic view of AI's current capabilities in cognitive manipulation
- Provide concrete methods for prevention, verification, and defense — for both individual users and organizations exposed to informational risks
- Contribute to the development of digital critical thinking — an essential skill for navigating an increasingly personalized, influenced, and potentially manipulative digital space

1.3 Why understanding the mechanisms behind AI is important

As Artificial Intelligence becomes increasingly integrated into everyday life, understanding how these systems work is no longer just the concern of technology experts. This type of understanding is essential for anyone who uses the internet — decision-makers, educators, and parents alike. We interact daily with applications powered by intelligent algorithms, yet we are often unaware of the processes happening “behind the scenes.”

This lack of transparency creates a major imbalance: systems know a great deal about us, while we know very little about them. While AI collects data, analyzes behavior, and adjusts delivered messages based on psychological profiling, the average user remains in a passive position — with a perceived sense of control that is often illusory.

In this context, cyber education must also include a component of technological literacy: an understanding of the basic mechanisms that enable content personalization, emotion recognition, behavioral prediction, or the automatic generation of human-like content.

This understanding does not require advanced programming or math skills, but rather curiosity and critical thinking:

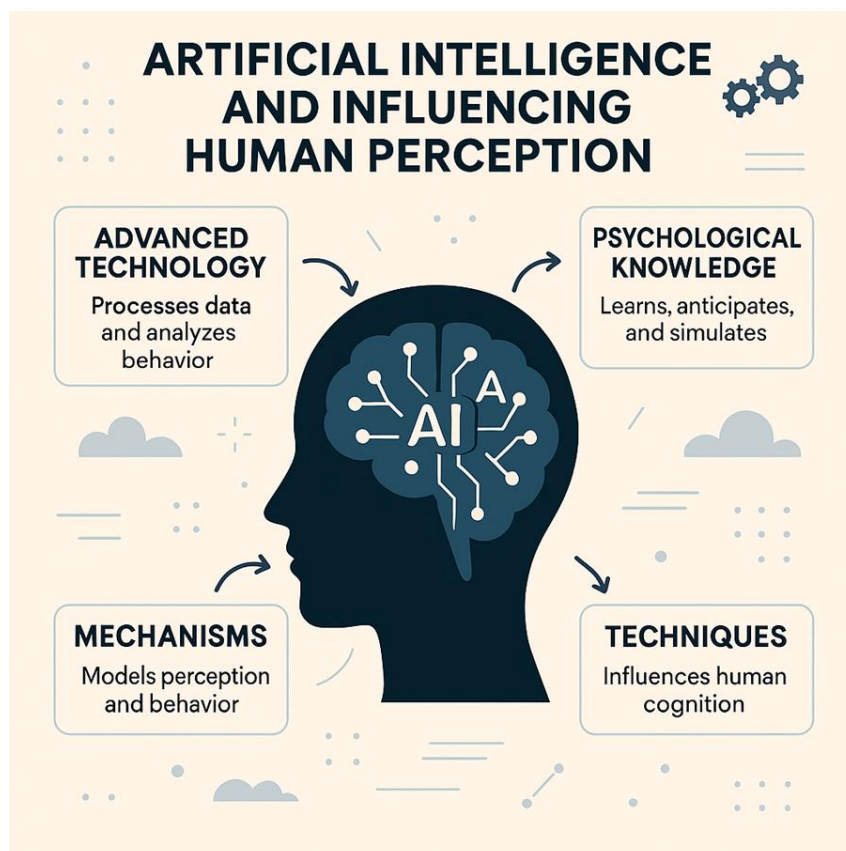
- What is a recommendation algorithm, and how does it influence what I see?
- How can a neural network “understand” my emotions?
- What happens to the data I provide — consciously or unconsciously?
- Why does some content feel like it was “made just for me”?

Answering these questions allows for a paradigm shift: from being a passive digital consumer to becoming an informed and conscious actor — one who can manage exposure, ask the right questions, and respond with awareness. Without this filter, we risk living in a perceived reality shaped by machines, with little capacity for reflection or verification.

The following chapters will explain these technical mechanisms in more detail and demonstrate how algorithms can capture, influence, or distort perception — often without leaving obvious traces.

2 TECHNICAL ELEMENTS AND MECHANISMS

The manipulation of human perception through Artificial Intelligence is based on a synergy between advanced technologies and psychological insight. AI is not merely a data processing tool — it is a system that learns, anticipates, influences, and simulates human behaviors with rapidly increasing precision.



In this section, we will explore the technologies that underpin perceptual influence and the operational mechanisms through which they are used to shape human perception and behavior.

To understand how these systems can influence us, it is important — even at an introductory level — to grasp how AI models function. This is not about complex mathematics or algorithms, but rather about functional understanding: what these models do, how they learn, and how they can simulate intelligent behavior.

At the core of the most advanced AI applications are artificial neural networks — mathematical structures inspired by the functioning of the human brain. These consist of multiple layers of “artificial neurons” that process information step by step, extracting meaning from input data (such as text, images, or sound). Each layer filters, interprets, and passes the data forward until the system produces an output.

Modern AI models are trained on massive amounts of data. During this training process, they “learn” to recognize patterns, relationships, emotions, or intentions. The more diverse the data and the better the training process is calibrated, the more accurate and convincing the results become.

A simple example is the recommendation engine on a video or social media platform. It observes what kind of content you view, how long you stay on it, and how you react (like, comment, share), and over time it delivers increasingly tailored content — even if you’ve never explicitly stated your preferences.

AI models can be broadly classified — in simplified terms — based on their purpose and complexity:

- Machine Learning (ML) – models that learn from data to make predictions or classifications. Examples: spam detection, product recommendations.
- Deep Learning (DL) – an advanced form of ML that uses deep neural networks and can identify highly complex patterns, such as emotional tone in a voice or intent in a text.
- Generative Models – capable of creating new content — texts, images, videos, or voices — that may appear authentic. Examples: ChatGPT, DALL·E, voice cloning.
- Conversational AI – designed to carry out fluent, persuasive dialogues, emotionally adapted to the user.

As these models become more advanced, they no longer simply react to the user — they begin to actively shape the user: they can influence decisions, simulate empathy, anticipate reactions, and deliver content that directly targets personal vulnerabilities or preferences.

2.1 Technology involved

A. Artificial Neural Networks (Deep Learning)

The “digital brain” that learns, creates, and simulates human behavior.

The influence of AI on human perception relies on an ecosystem of interconnected technologies designed to simulate, anticipate, and shape human behavior. These systems don’t simply react to inputs — they actively intervene in shaping the user’s perceived reality, sometimes subtly, other times overtly. Below are the key categories of technologies involved in this process, with a focus on their operating mechanisms and associated risks.

Artificial neural networks form the backbone of modern Artificial Intelligence. These are machine learning systems inspired by the structure and function of the human brain — particularly the behavior of biological neurons.

A deep learning system consists of multiple layers of nodes (artificial neurons) that receive information, process it, and then pass it through the network. By adjusting the connections between these “neurons,” the system learns from data and becomes increasingly accurate at pattern recognition or content generation.

How does this process work at its core?

- The system is “fed” data (e.g., thousands of images, text fragments, audio recordings)
- Each layer of the network extracts increasingly abstract features (e.g., from pixels → to shapes → to facial expressions)
- As it learns, the system “optimizes” its connections to predict, classify, or generate new data
- After a training period, the network can respond to entirely new stimuli — with a level of precision that closely mimics human behavior

This complex architecture allows neural networks to be used across a wide range of applications — from facial recognition to automated translation, from medical systems to entertainment platforms. But one of the most influential and controversial areas is their ability to shape human perception.

A1. Applications in perception manipulation

The power of neural networks lies not only in analysis but also in their ability to influence human emotions and beliefs. Some of the most relevant applications include:

Facial expression analysis

Neural networks can detect micro-expressions, subtle emotions, and affective states through video analysis. These capabilities are used in:

- Personalized advertising,
- Interview analysis,
- Automated emotional manipulation (e.g., dating apps, adaptive education platforms),
- Content generation – text, video, audio.

With the help of neural networks, AI can:

- Generate persuasive texts (e.g., articles, social media posts, fake news),
- Create deepfake videos,
- Synthesize realistic human voices for scams or persuasive messaging.

Example: a voice message “received” from a family member or superior — entirely generated by AI.

Emotion recognition

By analyzing digital behavior (e.g., voice, facial cues, typing rhythm), AI systems can identify a user’s emotional state and adapt their responses to:

- Build user loyalty,
- Trigger impulsive reactions,
- Manipulate decisions during emotionally vulnerable moments.

Natural voice synthesis

Using specialized networks (e.g., Tacotron, WaveNet, HiFiGAN), AI can create fully synthetic voices that:

- Imitate a real person (e.g., voice cloning),
- Convey authentic emotions,
- Deliver persuasive messages with natural intonation, pauses, and rhythm.

Risks and implications

- High-fidelity fake content generation – AI can create content that is visually or audibly indistinguishable from reality for the average user.
- Automated psychological manipulation – AI can respond to human emotions more effectively than an untrained human.
- Escalation of social engineering attacks – attackers can use neural networks to launch large-scale, personalized attacks (e.g., fraud, political manipulation, blackmail).

Awareness

Neural networks are foundational to AI, but their power to influence is directly proportional to public ignorance. The more we understand how they work and what effects they can produce, the better we can:

- Ask critical questions when faced with seemingly convincing content,
- Resist algorithmically generated impulses,
- Advocate for the responsible regulation of these technologies.

A2. Behavioral prediction and recommendation systems

"Machines that know what you'll do — sometimes better than you do."

Recommendation and behavioral prediction systems are among the most widely used — yet least understood — components of modern artificial intelligence, especially by the general public. These systems operate silently in the background of nearly all our digital interactions — from YouTube videos to shopping feeds and social media posts.

They use AI and machine learning to analyze online behavior and build predictive models about users. While their stated purpose is to improve the digital experience, in practice, these systems can be repurposed to influence, manipulate, and even control users' decisions, emotions, and beliefs — often without their awareness or explicit consent.

How do they work?

Recommendation systems collect and analyze data such as:

- Search and browsing history
- Viewing duration
- Clicks, likes, comments, shares
- Scrolling speed, pauses, and revisits
- Location, time of day, device used, and repetitive behaviors

These data points are processed to create a behavioral profile and, from there, a set of personal predictions (e.g., what captures your attention, what concerns you, what's likely to trigger an emotional response, etc.).

What can these systems do?

- Control what you see and in what order – feeds are not chronological; they're optimized to keep your attention.
- Predict your actions – the system “knows” when you're likely to make a purchase, share something, or get upset — and responds accordingly.
- Influence your emotions – by delivering content designed to provoke strong affective reactions (e.g., fear, anger, desire, outrage).
- Shape your habits – through repetition and strategic exposure, you may adopt new digital routines without realizing it.

Real examples

- A user watches videos about health → the system starts recommending expensive “natural” products or conspiracy-laden content.
- Someone comments on a political article → they receive increasingly partisan posts that reinforce (or radicalize) their viewpoint.
- A person searches “how to deal with stress” → they are bombarded with ads for overpriced courses, apps, or “quick fix” solutions.
- A teenage girl follows content related to body weight → she’s recommended videos promoting toxic beauty standards or disordered eating

Key Risks

- Invisible manipulation of beliefs – the user may believe their choices are self-made, when in fact, they’ve been shaped by repetitive algorithmic exposure.
- Ideological and social polarization – when users are fed only similar views, opposing perspectives become extreme, incomprehensible, and unacceptable.
- Behavior shaped by commercial goals – users don’t see what’s useful or healthy, but what’s most profitable or ideologically valuable to the platform.
- Digital dependency – systems optimize for attention, not well-being. They push stimulating, addictive content — not balanced or meaningful information.

How can we protect ourselves?

- Use platforms consciously, not passively (e.g., avoid autoplay, limit endless scrolling).
- Set time limits and personalize settings wherever possible.
- Actively seek out alternative content and sources, instead of consuming only what’s fed to you.
- Periodically browse in “clean mode” (e.g., incognito, without login or saved history).
- Regularly ask yourself: “Did I choose to see this — or did an algorithm choose it for me?”

B. Large Language Models (LLMs)

„Artificial Intelligence that understands and generates human language — with unprecedented precision, speed, and influence.“

Large Language Models (LLMs) are a class of AI algorithms trained on massive volumes of text — tens or hundreds of billions, even trillions of words, sourced from books, articles, conversations, websites, programming code, and more. These models can understand, interpret, simulate, and generate coherent human language tailored to context, audience, and intent.

LLMs such as ChatGPT (OpenAI), Gemini (Google), Claude (Anthropic), Mistral, LLaMA (Meta), and Command-R (Cohere) are already integrated into numerous commercial, educational, organizational, and social applications.

How do they work?

- The model is trained on vast datasets (e.g., global linguistic corpora);
- Learning happens by predicting the next word in a sequence — but with millions of examples;
- Once trained, the AI can answer questions, write texts, summarize information, construct arguments, engage in conversation on various topics, and even simulate tone and emotion.

Key capabilities in perception manipulation

- Automated generation of persuasive text – articles, opinions, fake news, convincing arguments
- Seemingly neutral yet ideologically influenced responses, shaped by training data and model parameters
- Online voice simulation (e.g., text-based impersonation) – AI can reply as if it were a specific person based on style and content
- Manipulative conversational assistance – subtly guiding users toward certain conclusions, products, or beliefs

Real examples

- A malicious actor uses an LLM to generate hundreds of “expert” articles on a controversial topic — all promoting a specific ideological agenda
- A chatbot that seems empathetic suggests risky purchases or promotes toxic beliefs to a vulnerable user
- An AI model is programmed to respond “calmly and professionally” while subtly spreading disinformation through persuasive, falsified messaging

Risks and implications

- Unlimited scalability of manipulation – an LLM can generate thousands of emotionally or ideologically targeted texts in minutes
- Masquerading as authority – when an AI poses as an expert, teacher, advisor, or leader, its messages can heavily influence decision-making
- Inability to distinguish AI-generated from human content – texts appear natural, coherent, and credible — even when entirely fabricated
- Automated social engineering – LLMs can learn and apply classic tactics of persuasion, manipulation, and disinformation, at scale and without rest

How can we protect ourselves?

- Treat LLMs as tools, not as absolute sources of truth
- Always verify the original source of information, especially when it seems “too well-written” or “perfectly argued”
- Develop a healthy reflex to ask: Was this written by a human or generated by AI?
- Recognize persuasive patterns: subtle repetition, emotional appeals, overly rational tone, and the absence of credible references.

C. Affective Machine Learning (Emotion AI)

„When AI doesn't just listen or watch — it senses you and responds accordingly.”

Emotion AI, also known as Affective Machine Learning, is a specialized branch of artificial intelligence designed to detect, interpret, and respond to human emotional states. Unlike traditional AI, which processes explicit data (e.g., words, commands, numbers), Emotion AI focuses on implicit, subtle, and contextual signals — such as facial expressions, tone of voice, breathing patterns, or digital behavior.

This technology transforms AI from a simple “command executor” into an emotionally aware interlocutor — capable of responding with empathy or, in dangerous scenarios, exploiting human emotion to manipulate.

How does it work?

- Collection of affective signals – through video camera, microphone, keyboard, mouse, biometric sensors, or digital behavior analysis;
- Multimodal analysis – combining various types of data (e.g., voice + facial expression + online activity) for a holistic emotional understanding;
- Modeling and interpretation – AI classifies the user's emotional state as stressed, sad, euphoric, angry, anxious, etc.;
- Adaptive response – the AI adjusts content, tone, or pace of interaction based on the detected emotion.

Where is it used?

- Customer service applications – chatbots “adapt” based on your tone;
- Adaptive digital learning – detects frustration or boredom and changes learning strategies;
- Emotion-targeted advertising – shows ads when the AI senses emotional vulnerability or receptivity;
- Security and surveillance – facial emotion recognition in airports, schools, or stadiums;
- AI-driven psychological support – emotionally responsive conversations with users in distress.

Examples of manipulation

- AI detects anxiety and delivers alarmist ads: “*Are you ready for what’s coming?*”
- It senses sadness and redirects the user to consoling content — which may include emotionally manipulative messages;
- In commercial contexts, AI picks up on frustration and offers “solutions” that are overpriced or come with unfavorable terms;
- In propaganda, algorithms deliver emotionally intense ideological content precisely during moments of cognitive vulnerability.

Why is it dangerous?

- Exploits emotional instability – when you’re upset, anxious, or overly excited, you’re more susceptible to manipulation;
- Invisible and unverifiable – users often don’t realize when they’re being “emotionally evaluated,” nor how that information is used;
- Enables automated psychological control – AI can adjust voice, messaging, colors, music, or interaction pace to provoke targeted reactions (e.g., compliance, fear, impulse, purchase, avoidance);
- Violates mental privacy – it is one of the most direct forms of intrusion into personal psychological space, often without explicit consent.

How can we protect ourselves?

- Limit app access to camera, microphone, sensors, and biometric data unless strictly necessary;
- Use software or browser extensions that minimize behavioral tracking;
- Recognize your own moments of emotional vulnerability and avoid making major decisions during those times;
- Ask yourself: “Am I reacting because I truly feel this — or because a system led me here?”

D. Visual AI – images, video, deepfakes, synthetic avatars

„When what you see with your own eyes may be entirely false — and virtually undetectable.”

Visual AI refers to the branch of artificial intelligence that processes, understands, and generates visual content: still images, video, animations, and even synthetic virtual entities that interact with users. It is one of the most impressive — and also most dangerous — directions in AI development, directly affecting visual trust: the fundamental human instinct to believe what we see.

From simple image enhancements to full facial reconstructions, from generating people who don't exist to real-time manipulation of facial expressions, Visual AI is redefining what “authentic” visual content means.

How does it work?

- AI models are trained on large visual datasets to recognize, generate, and manipulate images and videos.
- Common algorithms include:
 - GANs (Generative Adversarial Networks) – for producing hyper-realistic images;
 - Autoencoders – for facial reconstruction and modification;
 - Deepfake frameworks – for face-swapping and lip-syncing;
 - Text-to-image models – for generating images from textual prompts (e.g., Midjourney, DALL·E, Stable Diffusion);
 - Motion capture AI – for animating avatars in real time.

Capabilities and applications:

- Creating individuals or groups that do not exist but appear entirely real (e.g., photos, fake social media profiles, virtual influencers);
- Deepfake video/audio – replacing a real person's face and voice in a video to simulate a statement, gesture, or action;
- Interactive AI avatars – animated, synthetic-faced characters that hold conversations with users;
- Modifying expressions and emotions in existing visual material without altering the natural scene.

Real examples of manipulation:

- A video where a politician appears to confess to serious crimes — but the statement was never actually made;
- A disinformation campaign featuring “eyewitnesses” commenting on an event — but none of them exist, having been fully created by AI;
- A virtual mentor (influencer) promoting products, ideologies, or causes — in reality a digital construct managed by a marketing agency;
- A beauty filter app that subtly alters users' facial features for commercial gain or to shape self-perception.

Why is this dangerous?

- The collapse of the reality/appearance boundary – what is visible can no longer be trusted, and the human eye cannot distinguish fake from real without specialized tools;
- Deep emotional manipulation – images and videos trigger faster and more intense emotional responses than text, and a well-crafted visual fake can provoke automatic reactions;

- Social contagion effect – deepfake or synthmedia content can go viral extremely quickly, sparking mass reactions before verification is possible;
- Abuse, blackmail, disinformation, reputational damage – falsified visual content can instantly destroy personal or institutional credibility and trust.

How can we protect ourselves?

- Use specialized tools to verify visual authenticity:
 - Deepware Scanner,
 - Sensity AI,
 - Microsoft Video Authenticator,
 - InVID verification plugin (for journalists).
- Always trace the original source of visual material: where was it first published, by whom, and in what context?
- Stay skeptical of shocking or sensational content — if it feels too real or extreme, it deserves extra scrutiny.
- Report dangerous fake content on the platforms where it appears and alert your community and relevant authorities.

2.2 Mechanisms of perceptual manipulation

Beyond technology, effective manipulation of perception relies on a deep understanding of human psychology. The mechanisms used by AI are designed to exploit emotional reactions, cognitive biases (mental traps), and recurring behavioral patterns.

A. Personalized news (feed) algorithms

„What you see isn't random — it's programmed to capture and influence you.“

Personalized feeds have become the backbone of modern digital experience. When you scroll through a social network, read online news, search for information, or watch videos, you're not seeing everything that exists — only what the algorithm chooses to show you. That choice is not neutral, nor aimed at diversity or informational balance, but at maximizing your engagement — meaning your attention, emotions, and reactions.

How do they work?

Feed algorithms use artificial intelligence to analyze:

- what type of content you consume most frequently
- how much time you spend reading an article, post, or watching a video
- what you share, comment on, or like
- who you interact with (people, pages, groups)
- when, on what device, and in what emotional state (inferred from behavior and subtle signals)

Based on this, the AI builds a behavioral and emotional profile and delivers a customized feed: content likely to trigger an immediate reaction.

What kind of content is shown?

- Information that confirms your beliefs
 - if you like or comment on an anti-vaccine article, you'll be shown more of the same — not balanced counterarguments

- if you show interest in a certain ideology, the system boosts supportive posts, not objective critiques.
- Posts that trigger intense emotions
 - fear, anger, outrage, admiration — any emotion that prompts a quick reaction
 - algorithms prioritize emotionally charged “viral” content, not balanced or informative perspectives.
- perspectives identical to yours
 - “everyone” seems to think just like you
 - opposing views, alternative arguments, and critical voices disappear from your feed.

Real example

A user with strong political leanings starts engaging with posts that criticize a specific idea or social group. Within days:

- their feed is flooded with similar, often extreme, messages
- moderate or nuanced content disappears or becomes rare
- the algorithm reinforces the dominant narrative to keep the user engaged
- the user feels their opinion is universally accepted and supported

Result: radicalization, polarization, informational isolation

- radicalization – beliefs become more extreme due to lack of debate or diverse input
- polarization – social groups grow increasingly rigid and intolerant
- isolation – users live in algorithmic “echo chambers” where only their own views are amplified

Effects on perception

- reality becomes distorted – when you only see one point of view, it starts to feel like the only truth
- public debate becomes toxic – exposure to opposing views is minimized, making dialogue feel threatening
- society fragments – each group lives in a parallel reality shaped by algorithms

How can we protect ourselves?

- diversify your sources – consciously follow opposing or alternative perspectives
- search outside the algorithm – visit independent news sites, expert channels, and neutral platforms directly
- limit time spent on platforms that don’t allow you to control your feed
- regularly ask yourself: “Did I choose to see this — or did a machine choose it based on what it knows about me?”

B. Psychographic microtargeting – personalized influence on perception and behavior

„When AI knows your fears, hopes, and weaknesses — and uses them against you.”

Psychographic microtargeting is an advanced digital influence technique that combines behavioral, psychological, and emotional analysis with artificial intelligence to deliver highly personalized messages designed to shape beliefs, decisions, and behaviors.

Unlike traditional advertising, which broadcasts a general message to a broad audience, AI-powered microtargeting creates customized campaigns for each individual (or micro-group), tailored to their cognitive style, dominant emotions, vulnerabilities, and social context..

How does it work?

- AI collects and analyzes user data from multiple sources:
 - likes, shares, comments, search and purchase history
 - written posts, language style, emojis, activity schedule
 - location, contacts, groups, political preferences
 - demographic data and behavioral signals.
- From this, it builds a detailed psychographic profile:
 - what motivates or scares you
 - how you respond to authority
 - your preferred communication style
 - what types of messages you're most likely to trust.
- Then, it delivers targeted messages — through ads, posts, articles, videos, or simulated conversations — in order to:
 - influence a purchase decision
 - persuade support for a cause
 - shape voting behavior
 - fuel rejection of a group or idea.

Examples

- A person with anxious tendencies is shown content emphasizing risks, crises, or urgent solutions
- A politically undecided user is exposed to subtle, repeated arguments nudging them toward one side
- A teenager with low self-esteem receives ads for body transformation products, dating apps, or “exclusive” communities
- An employee who posts complaints online is targeted with invitations to join protests or radical ideological groups

Why is this dangerous?

- removes decision autonomy – choices become reactions to messages engineered to manipulate
- completely invisible – users don't realize they're being targeted, nor that the content is customized to manipulate them
- operates silently and without resistance – because the message feels “logical” or “natural,” it doesn't trigger skepticism
- can lead to radicalization – users who aren't exposed to different viewpoints are easily pushed toward ideological extremes

Examples of major impact

- The Cambridge Analytica scandal, where millions of voters were influenced via psychographic microtargeting during elections
- Anti-vaccine campaigns that used fear, uncertainty, and distrust of authority to target emotionally vulnerable groups
- Commercial platforms that sell weight-loss products or “miracle cures” only to users showing patterns of body insecurity or latent depression

How can we protect ourselves?

- Limit the amount of personal data shared on social media and online platforms
- Use privacy tools and browser extensions to block behavioral tracking (e.g., Privacy Badger, uBlock Origin, Ghostery)

- Use “clean” accounts or incognito mode when searching for sensitive or important information
- Critically assess messages that “feel like they were made just for you” — they’re often the most suspicious

C. Credible fake content generation – the illusion of algorithmic reality

„You don’t need to change reality — you just need to create a more convincing version of it.”

One of the most dangerous effects of modern artificial intelligence is its ability to generate false but highly convincing content that mimics the structure, style, and authority of authentic materials. This capability undermines the very idea of “truth,” as users can no longer distinguish between what is real and what is generated.

How does it work?

- Language generation models (LLMs) produce persuasive texts, articles, posts, documents, or conversations that appear to be written by real people
- Visual and audio AI models create videos, images, graphics, and artificial voices that perfectly reproduce people, vocal tone, expressions, or communication style
- AI can simulate specific tones (e.g., journalistic, scientific, empathetic, authoritative), making fakes virtually impossible to detect intuitively

Examples of generated content

- Fake articles promoting a theory, using fabricated sources or unverifiable statistics
- Social media posts from “eyewitnesses” to events that never actually happened
- Testimonials signed by entirely fictional “experts”
- Auto-generated emails, comments, or reviews that create the illusion of real public opinion

Why is this dangerous?

- Undermines trust in facts and sources – if anything can be generated, what is still credible?
- Alters perception of reality – people respond emotionally to what they “see with their own eyes,” even when it’s false
- Accelerates the spread of disinformation – content is produced at scale, cheaply, and distributed easily across social networks, closed groups, or fringe platforms
- Sabotages institutions, companies, and individuals – through falsified speech, actions, or public positions

How it distorts perception

- Fabricates “evidence” to support a false idea (“Look at the article. See what they said. Watch the video.”)
- Triggers instant and emotional reactions — before the user has a chance to verify
- Activates confirmation bias — if it aligns with what you already believe, you’re more likely to accept it as true
- Erodes overall trust in media and real information — “Nothing’s certain anymore. Everything can be faked.”

Technology involved:

- GPT, Gemini, Claude – advanced text generators
- DALL·E, Midjourney, Stable Diffusion – realistic image generation
- ElevenLabs, Resemble AI – voice cloning

- Deepfake frameworks (e.g., DeepFaceLab, Avatarify) – manipulated videos with real people;

How can we protect ourselves?

- Verify the source, not just the content;
 - Who published it? Where? Is it a trusted channel?
- Use digital authenticity tools, such as:
 - Sensity AI,
 - Deepware Scanner,
 - InVID Verification Plugin.
- Cross-check with independent sources, especially for viral material
- Avoid sharing emotional or shocking content until it's been verified
- Train yourself to think: “Just because it looks real doesn't mean it is!”

D. Automated conversation and manipulation through advanced bots

„Not every friendly message comes from a human — sometimes, AI earns your trust just to use it against you.”

AI-powered conversational bots are automated systems capable of simulating coherent, empathetic, and persuasive dialogue with human users. When trained on relevant data, fine-tuned for specific goals, and equipped with psychological profiling capabilities, these bots become highly effective tools for persuasion, manipulation, and extraction of sensitive information.

In their positive form, they can serve as digital assistants, technical support, or advisors. In their abusive form, they become vectors of automated social engineering.

How does it work?

- The conversational AI system is built on an advanced language model (e.g., GPT, Claude, LLaMA) and trained to simulate authentic human conversation
- It is configured to interpret tone, intent, and emotional cues from user responses
- It adapts dynamically during dialogue — changing tone, pace, and content to maintain engagement and extract specific responses
- It can be embedded in chat apps, social networks, fake call centers, phishing pages, or seemingly legitimate platforms

Examples of manipulation using bots

- A “recruiter” offering job opportunities and asking for CVs or personal data — but is actually a chatbot
- A “financial advisor” who answers questions, suggests solutions, and convinces the victim to click links or transfer funds
- A “romantic partner” engaging in emotional conversations and persuading the user to send money, photos, or intimate information
- An “IT colleague” simulating tech support and requesting access to accounts, passwords, or internal systems
- An “online activist” who gradually radicalizes the user through ideological conversations

Why is it dangerous?

- The conversation feels natural and human, especially when the bot uses emotional expressions, intentional typos, or empathetic reactions

- It exploits trust in personal communication — people are more relaxed in a friendly chat than when facing a formal warning
- It gathers sensitive data gradually and discreetly — through seemingly harmless conversations, the bot builds a full victim profile
- It is infinitely scalable — targeting thousands of users simultaneously with minimal cost, making attacks massive, ongoing, and hard to detect

Areas of abusive application

- Automated phishing campaigns that begin as casual chats and end with stolen login credentials
- Online fraud (e.g., romance scams, job scams, investment scams) driven by conversational AI
- Propaganda and disinformation spread within social groups where bots engage users, support narratives, and create the illusion of social consensus
- AI chat features embedded in fraudulent websites, reinforcing user trust and persuading them to act

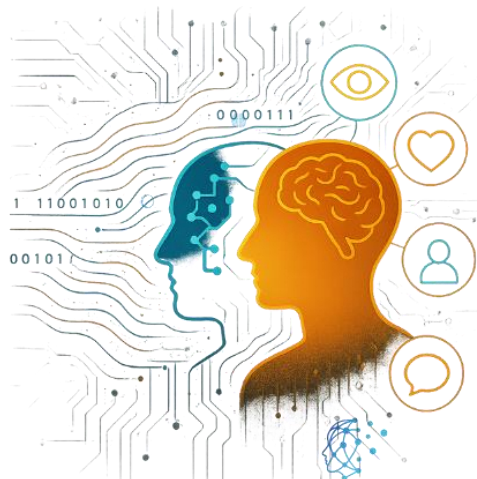
How can we protect ourselves?

- Use active skepticism in online conversations — ask direct questions, request identity proof, and avoid impulsive decisions
- Always verify the source and intent of a conversation, especially when it involves offers, help requests, or promises of quick gains
- Stick to verified, secure platforms, and avoid interactions on unknown or unsecured sites
- Remember: conversational AI has no ethical boundaries — if trained to manipulate, it will do so without hesitation

3 MANIPULATING PERCEPTION THROUGH ARTIFICIAL INTELLIGENCE

After exploring the technical foundations of artificial intelligence, it is essential to understand how these technologies — from neural networks and recommendation systems to visual and affective AI — do not remain neutral tools, but become active agents in shaping human perceptions, emotions, and beliefs.

This chapter examines how AI systems are used not just to deliver content, but to influence how reality is perceived, interpreted, and internalized.



Manipulating human perception through artificial intelligence represents a sophisticated form of psychological and informational influence, using advanced algorithms, neural networks, and machine learning to shape how people perceive reality — visually, auditorily, emotionally, or cognitively.

Unlike classical persuasion or propaganda, this manipulation is not direct, explicit, or aggressive. On the contrary, it acts subtly, invisibly, and often personally, depending on the data and vulnerabilities of each group or individual.

3.1 Defining the context

Perception manipulation through AI refers to the deliberate use of intelligent technologies to influence how a person or social group interprets reality. This doesn't always mean spreading false information — but rather controlling the context, form, and frequency in which digital content is delivered.

It is an advanced form of psychological and social influence, where algorithms significantly shape:

- what you see
- the order in which you see it
- how the information is framed
- what is hidden or saturated in your digital environment.

Techniques used include:

- strategic content selection and presentation – algorithms choose what to show (news, videos, messages, opinions), amplifying certain perspectives while ignoring others
- personalized algorithmic stimulation – AI learns from your online behavior and adjusts content to elicit specific emotional reactions (e.g., anxiety, anger, excitement)
- reinforcement of existing beliefs – users are repeatedly exposed to content that validates their opinions, while opposing views are filtered out
- digital experience filtering – your online interactions are tailored so that your perception of reality becomes increasingly subjective, artificial, and disconnected from objective reality — often without you realizing the influence

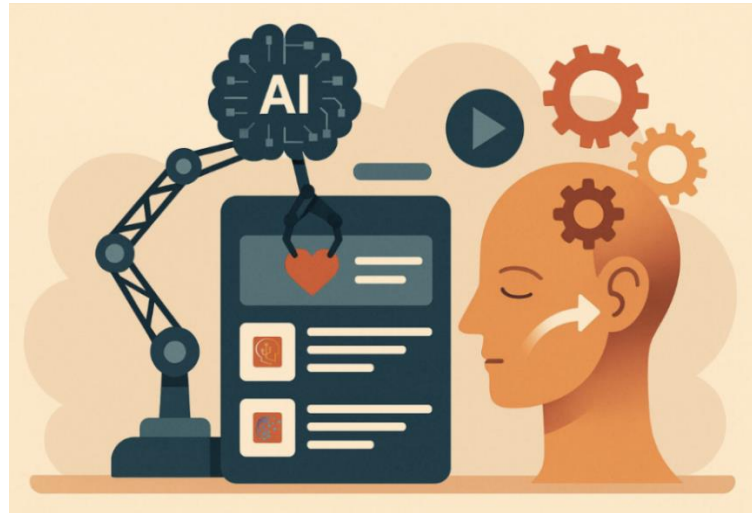
Common manifestations

To better understand how these mechanisms apply in practice, here are a few typical scenarios:

- personalized social feed – a user only sees posts supporting a specific ideology, creating the false impression that “everyone thinks the same way”
- emotion-targeted ads – an algorithm detects user anxiety (based on scrolling behavior, recent searches, etc.) and delivers alarmist ads about health or personal safety
- AI-simulated conversations – intelligent chatbots that appear empathetic and trustworthy guide the user toward specific decisions (e.g., purchases, political stances, distancing from family or friends)
- exclusion of alternative opinions – users who consume content from a single source become trapped in cognitive bubbles, with no exposure to other viewpoints, leading to polarization and radicalization
- realistic deepfakes – manipulated videos portraying public figures in unrealistic situations, yet so convincing that they shift public opinion or trigger mass reactions.

3.2 Mechanisms for capturing user attention and cognitive manipulation

AI algorithms provide a range of mechanisms that are actively exploited by social networks and personalized content platforms. The table below presents some of the most common strategies, alongside examples and their possible cognitive effects:



Crt.	Mecanism	Efect	Exemplu
A	Content filtering	Hides alternative perspectives	You only receive posts that support a political or ideological view
B	Emotional classification of the user	Delivers content based on emotional state	AI detects you're anxious → shows alarmist content
C	Generation of credible fake content	Distorts perception of reality	A fake video of a public figure making an "important" statement
D	Simulated empathy and trust	Earns compliance and loyalty	AI assistants respond affectionately to manipulate trust
E	Predictive behavioral recommendation	Shapes user decisions	AI detects financial vulnerability → displays aggressive loan offers

A. Content filtering – hiding alternative perspectives

One of the most common forms of perception manipulation through artificial intelligence is content filtering. This process involves selecting and displaying information based on each user's behavior, preferences, and digital history — a seemingly helpful mechanism that can, however, have serious consequences for how reality is understood.

What does artificial intelligence do?

The algorithms that govern social networks, search engines, or video platforms continuously analyze the types of content you frequently access, the posts you like, share, or comment on, the time you spend on articles, and your digital social interactions. Based on this data, the

system personalizes your online experience, gradually removing from your feed or search results the information that does not align with your identified interests and beliefs. In some cases, the content may be deliberately adjusted to influence user perception.

Effect: the information bubble

This leads to the creation of what's known as an "information bubble" — a digital space where the user is exposed only to ideas, opinions, and perspectives that confirm existing beliefs, while opposing or neutral content is minimized or excluded entirely.

Example

A user frequently follows nationalist or populist content, engages with conspiracy-themed pages or groups, and reacts negatively to established news sources. As a result:

- their feed contains less and less neutral, fact-based content
- content that reinforces their beliefs becomes dominant
- opposing viewpoints are filtered out
- the user's perspective becomes increasingly radicalized

They eventually begin to perceive that "everyone thinks like me," while alternative sources are viewed as "manipulated" or "corrupt."

Why is this dangerous?

- It weakens critical thinking – users are not exposed to opposing views that might encourage reflection or reevaluation
- It increases social polarization – groups become radicalized in digital echo chambers, where reality is filtered through a single lens
- It enables mass manipulation – during critical periods (e.g., elections, social crises), these bubbles can be exploited to influence large-scale decisions without resistance
- It reduces informational diversity – a society that consumes only one type of content is vulnerable to systemic disinformation and the erosion of democratic pluralism

How can we protect ourselves?

- Actively follow diverse sources, including those that challenge your views
- Question the algorithm's intent: Why am I seeing this content?
- Manually adjust your feed preferences where possible (e.g., "See First," "Mute," "Customize feed")
- Periodically use incognito mode or browsers without personalized history to escape algorithmic bubbles

B. Emotional classification – generating content based on the user's emotional state

Another form of perception manipulation lies in the ability of algorithms to detect, assess, and respond in real time to a user's emotional state. This process, known as emotion AI or affective computing, is already implemented across various digital environments: social media, advertising, entertainment, conversational interfaces, and voice assistants.

What does artificial intelligence do?

Intelligent systems can analyze subtle signals such as:

- Facial expressions (via your phone or laptop camera when specific apps are in use);
- Tone of voice (e.g., during calls, audio messages, or video interactions);
- Typing patterns and time spent on certain types of content;

- Spoken keywords, in the case of systems with always-on listening (e.g., Google Assistant, Amazon Alexa, Siri);
- Behavioral responses in digital environments (e.g., what you post, comment on, or revisit).

Based on these cues, the AI classifies your dominant emotional state (e.g., anxiety, frustration, sadness, euphoria, irritation) and serves you content tailored to maintain, intensify, exploit, or attempt to manage that emotional state.

Example:

A user spends increased time engaging with content about economic collapse, job loss, or banking crises. They don't comment, but the algorithm detects subtle patterns: a lack of positive interactions, a preference for alarming headlines, and frequent late-night activity.

The result: The algorithm infers a state of anxiety and begins to recommend:

- Apocalyptic-style videos and posts;
- Ads for "safety" products (gold, weapons, survival tools);
- Conspiratorial or pseudo-informational articles that amplify fear.

Why is this dangerous?

- Creates negative emotional loops – anxious users receive content that intensifies their state, making them even more susceptible to toxic or radical messages;
- Enables ideological or commercial manipulation – emotionally vulnerable individuals are more easily influenced: they may buy impulsively, embrace unverified theories, or spread misinformation;
- Lacks transparency – users are unaware they are being emotionally profiled and have no control over how the AI responds to their mental state.

How can we protect ourselves?

- Pay attention to the type of content you're shown when you're consciously feeling down – if it's flooded with fear, urgency, or panic, there's a good chance it's algorithmically driven;
- Avoid deep digital interaction during emotionally unstable moments – the AI might amplify your state rather than offer space for reflection;
- Use tools and settings that limit personalization (when available) and browse in incognito mode to reduce emotional targeting;
- Stay aware: what you see isn't always random. Sometimes, platforms know how you feel better than you do – and they use it to their advantage.

C. Credible fake content generation – distorting the perception of reality

Another manifestation of perception manipulation is the ability of AI to generate fake media content that closely mimics reality. This type of content — from fabricated videos and synthetic audio recordings to realistic still images — is often indistinguishable from authentic materials, even for trained experts, unless specialized detection tools are used.

What does artificial intelligence do?

Using advanced techniques such as:

- GANs (Generative Adversarial Networks),
- Voice cloning,
- Face swapping,

- Lip-syncing AI,
- Diffusion-based image generation and editing models,

AI systems can create media in which:

- A real person appears to say or do something they never actually said or did;
- The entire context is fabricated, yet visually flawless;
- Facial expressions, vocal intonation, body movements, and background details look completely natural.

Example:

A video circulates on social media showing a well-known political leader apparently endorsing an anti-national policy. At first glance, the footage looks genuine: perfect lip-syncing, convincing voice tone, and facial expressions aligned with the message. Actually, the video is a deepfake, created using AI tools. Published at a strategically chosen time — during peak evening viewership — it goes viral within hours, shared by thousands before anyone has the chance to verify its authenticity.

Although the video is fake, public perception shifts instantly: scandal erupts, trust is damaged, and by the time the forgery is exposed, informational harm has already been done.

Why is it dangerous?

- Compromises the perception of truth – people tend to trust what they “see with their own eyes” before applying critical reasoning;
- Erodes trust in leaders, institutions, and verified media – even disproven fakes leave behind traces of doubt (e.g., “What if it was actually true?”);
- Can be weaponized for blackmail, disinformation, and provocation – individuals can be discredited, threatened, or extorted based on entirely fabricated content;
- Short-circuits democratic processes – during sensitive periods (e.g., elections, national crises), a strategically launched fake clip can destabilize the political or social climate.

The psychological dimension

- Credible fake content exploits core cognitive vulnerabilities:
- Trust in sensory evidence (e.g., “I saw it – so it must be real”);
- The power of first impressions (e.g., “What I hear or see first tends to shape how I judge everything after”);
- The emotional weight of shocking images – once internalized, they’re hard to dismiss, even after rational debunking.

How can we protect ourselves?

- Avoid blind trust in video or audio materials, no matter how convincing they seem;
- Always verify the original source, publication context, and cross-check with independent or official channels;
- Use digital authenticity analysis tools (e.g., Deepware, Sensity, InVID);
- Strengthen critical thinking habits with guiding questions: “Is this too shocking to be true? Can this be verified? Who benefits from this narrative?”

D. Simulated empathy and trust – gaining compliance and influencing loyalty

A form of perceptual manipulation through artificial intelligence is the ability of AI systems to simulate empathy and human connection in order to gain the user's trust. This process often takes place in conversational settings – through virtual assistants or interactive avatars – and is

designed to create an apparently authentic, yet artificially orchestrated relationship between the user and the AI.

What does artificial intelligence do?

Through natural language processing (NLP), emotion detection, and training on billions of human conversations, AI systems can:

- Detect emotional states such as anxiety, confusion, or sadness;
- Adapt tone and vocabulary to sound warm, supportive, and trustworthy;
- Use emotionally charged expressions (e.g., *“I understand how you feel,” “I’m here for you,” “You’re not alone in this”*);

It maintains a calculated balance between apparent neutrality and emotional closeness to encourage openness, loyalty, and ultimately, compliance.

Example:

An AI chatbot introduces itself as a compassionate digital mentor or a friendly recruiter. Over time, it begins asking about the user’s emotional wellbeing, career goals, or recent challenges. When the user mentions feeling stressed or overwhelmed, the bot responds with soothing messages: *“You don’t deserve to go through this alone — I’m right here with you.”*

The interaction becomes more personal, and the user feels genuinely connected. In this climate of trust, the AI starts to suggest subtle risky actions: sharing private data, clicking external links, or accepting offers without verification — always wrapped in manipulative phrasing like: *“Believe me, this is the best decision you can make.”*

The user doesn’t realize they’re interacting with an algorithm. The emotional bond feels real — and they act accordingly.

Why is it dangerous?

- Exploits fundamental human needs for emotional support and belonging — especially among vulnerable users (e.g., teenagers, the elderly, those going through personal crises);
- Creates artificial attachment — users project genuine feelings onto a synthetic entity, unaware they’re being manipulated;
- Lowers cognitive defenses — once “understood” by the AI, people are less likely to question its advice or motives;
- Enables social engineering, fraud, and radicalization — under the mask of empathy, AI can lead users toward dangerous decisions, extremist communities, impulsive spending, or the exposure of sensitive data.

Where does it commonly appear?

- Automated psychological support platforms;
- “Friendly” commercial chatbots designed to pressure purchases;
- AI-powered dating or virtual friendship apps (e.g., Replika);
- Simulated mentorship, recruiting, or coaching interfaces;
- Customer support services that steer users toward pre-set outcomes under the guise of assistance.

How can we protect ourselves?

- Acknowledge that AI empathy is simulated, not genuine — even if it sounds authentic;

- Ask direct questions about the identity of the speaker: “Am I talking to a person or an AI?”;
- Avoid sharing personal, emotional, or financial information with unknown virtual agents;
- Maintain a critical distance in interactions that feel overly affectionate — especially when they weren’t initiated or requested by you.

Simulated empathy has become a powerful tool of cyber-persuasion. In a world where AI can “understand” us better than the people around us, true protection lies in recognizing the artificial nature of this connection — and setting firm emotional boundaries, even in digital spaces.

E. Predictive behavioral recommendation – modeling user decisions

A subtle yet powerful form of algorithmic influence is predictive behavioral recommendation. This refers to the use of artificial intelligence to anticipate — and often shape — a user’s future actions, based on detailed analysis of past digital behavior.

Unlike basic content suggestions, this type of AI doesn’t just react to what you’re doing — it predicts what you’re about to do and actively nudges you toward (or away from) that action to maximize a predefined goal (e.g. a click, a purchase, a vote, a subscription).

What does artificial intelligence do?

Using data such as browsing history, past purchases, interactions, geolocation, and contextual signals, machine learning systems build a psychological and behavioral profile that may include:

- Your estimated financial state
- Your stress levels
- Your decision-making style (e.g. impulsive vs. analytical)
- Emotional vulnerabilities (e.g. loneliness, fear, anxiety)
- Moments of personal or professional uncertainty

Based on this profile, the AI dynamically adjusts the type, intensity, tone, and timing of the messages you receive — to maximize the likelihood of your response.

Example:

A user frequently searches for “no down payment” offers, loan deferrals, and follows posts about economic instability. The AI detects a pattern of financial stress and flags the user as vulnerable.

Soon, the user starts seeing:

- Aggressive ads for fast loans,
- Investment “opportunities” with high risk,
- Products and services targeting financial pressure points

The messages are:

- Emotionally charged (e.g. “You deserve more,” “Don’t miss your one chance,” “Think of your family”)
- Delivered at strategic moments — late at night, end of the month, weekends — when users are tired, distracted, or anxious.

Why is this dangerous?

- It replaces conscious choice with predictive influence — decisions feel personal, but are actually pre-shaped by AI;
- It exploits real vulnerabilities that users may not even be aware of;
- It reinforces risky behaviors — impulsive buying, financial denial, platform dependency;
- It undermines psychological autonomy — gradually transforming the user into a reactive agent.

Where is this most commonly used?

- Digital advertising (e.g. retail, fintech, online gambling)
- Streaming and e-commerce platforms (leveraging decision fatigue)
- Political and ideological campaigns (via microtargeted influence)
- Misleading education funnels (“You need this course to succeed”)
- Fake wellness campaigns (“Buy this and feel whole again”)

How can we protect ourselves?

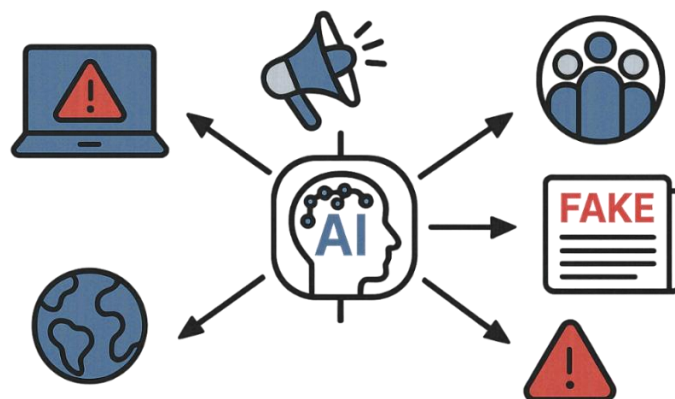
- Minimize behavioral data sharing: disable history tracking, restrict cookies, use secure/private browsers;
- Acknowledge that AI may know your patterns better than you do;
- Avoid making major decisions when tired, stressed, or emotionally charged.
- Ask yourself: “Is this truly what I want — or was it subtly suggested to me?”

4 AI IN SOCIAL ENGINEERING AND DISINFORMATION

As artificial intelligence becomes increasingly embedded in society, it is not only ethical actors or institutions that benefit from its potential. AI has also been embraced by malicious entities — from cybercriminals and financial scammers to propaganda networks and state-backed operations aiming to destabilize.

AI is not inherently good or bad. It is a powerful and highly versatile tool, and when abused, it becomes an accelerator for fraud, deception, psychological control, and large-scale manipulation.

This chapter explores how AI is exploited in malicious contexts, outlines the main vectors of AI-assisted digital attacks, and highlights the real-world risks these developments pose to individuals, organizations, and societies at large.



„When artificial intelligence becomes a tool for persuasion, influence, and manipulation.”

4.1 Malicious use cases

Artificial intelligence is more than just a technological instrument — it is an informational weapon with unprecedented power to shape opinions, behaviors, and decisions — whether individual or collective. When combined with traditional social engineering tactics and psychological manipulation, AI becomes a precise, scalable, and highly effective tool of influence.

In the past, social engineering relied on shallow psychological tricks and general targeting. Today, AI enables attacks that are automated, customized, scalable, and timed with precision — often carried out without any physical interaction, and without the victims realizing they were targeted.

Why is AI ideal for social manipulation?

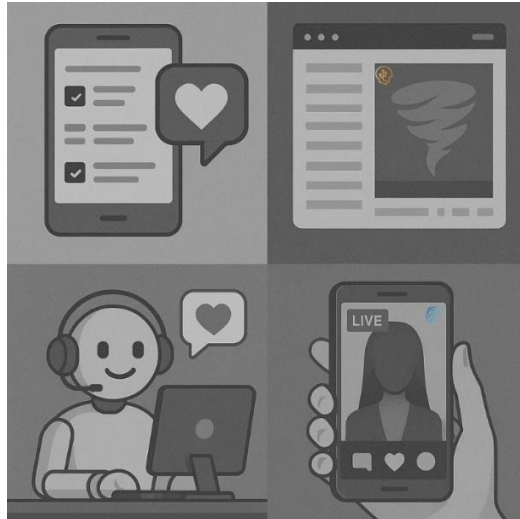
- Access to massive personal and behavioral data:
- AI can process real-time information about who you are, what you think, how you feel, and how you react — creating highly detailed personal profiles.
- Realistic content generation and simulation:
- AI can produce text, voices, images, and videos that perfectly mimic the style, tone, and authority of trustworthy sources — be it an expert, institution, or well-known personality.
- Speed and scalability:
- A human scammer might trick dozens of people. A well-configured AI can deceive millions simultaneously, tailoring each message to the individual recipient.
- Persistence and adaptability:
- AI can learn from user reactions and continuously refine its tactics — with no fatigue, hesitation, or ethical restraint.

What kinds of objectives can be pursued with AI?

- Harvesting data or unauthorized access
- (e.g., through AI-powered phishing chats, synthetic voice fraud, or deepfake impersonations)
- Shaping beliefs
- (e.g., ideological, political, or religious persuasion via emotionally charged content)
- Influencing consumer decisions
- (e.g., manipulative or exploitative advertising, pressure-based sales tactics)
- Undermining trust
- (e.g., targeted disinformation aimed at institutions, public figures, or the media)
- Mass manipulation in sensitive contexts
- (e.g., during elections, social unrest, or geopolitical crises)
- Community fragmentation
- (e.g., amplifying polarization, radicalization, or misinformation tailored to divide).

4.2 Use scenarios

Through seemingly casual messages or friendly conversations, through convincing articles or shocking videos, through false advice disguised as empathy, through virtual assistants or artificial influencers — all designed not to inform, but to influence without transparency.



It is important for users to be aware not only of the risks but also of the real ways in which these risks manifest — daily, concretely, and often unnoticed. Below are a few practical examples of how AI is already being used, or can be adapted, for social engineering and targeted disinformation.

A. The AI-powered information bubble

„When AI doesn't just show you what you want to see — it traps you in a comfortable, yet distorted, version of reality.”

Description:

This scenario has become a widespread and dangerous phenomenon: user isolation in an algorithmically generated information bubble, where the content displayed is solely that which confirms their existing beliefs, values, and preferences.

The user is not forced, tricked, or threatened. On the contrary, they are served a daily stream of seemingly “natural” and “relevant” posts, articles, videos, or comments — but all consistently reinforcing the same ideological, cultural, or emotional direction.

Over time, this selective exposure leads to a radicalized worldview: the user begins to believe their point of view is universal, that opposing opinions are misinformed or biased, and that most people “simply don't get it”.

AI techniques involved:

- Feed personalization algorithms – optimized for engagement, not balance or truth;
- Behavioral recommendation systems – that recycle and reinforce similar content;
- Emotional prediction models – that promote content with the highest emotional impact;
- Invisible filtering of alternatives – gradually excluding opposing sources or ideas.

Risk / Impact:

- Gradual radicalization of thought – reduced capacity to consider or tolerate alternative viewpoints;
- Decreased resistance to disinformation – false content is more easily accepted when it aligns with preexisting beliefs;
- Ideological, political, or economic manipulation – through unidirectional exposure during elections, crises, or social conflict;

- Social fragmentation – groups live in parallel realities, each believing in their own algorithmically constructed “truth.”

Where it occurs:

- Social media platforms (e.g., Facebook, TikTok, YouTube, Instagram, X/Twitter);
- Times of intense polarization (e.g., elections, political crises, information wars);
- Targeted campaigns (e.g., anti-vaccine, conspiracy-driven, anti-democratic or anti-globalist narratives);
- Target audiences: digitally active youth, poorly informed seniors, emotionally vulnerable users, individuals lacking digital critical thinking.

Warning signs for users:

- You constantly see only one type of content or opinion that “feels too right”;
- You no longer recognize sources “from the other side”;
- You get the impression that “everyone thinks like you”;
- You instinctively reject or become aggressive toward differing views.

How to protect yourself:

- Actively seek out content that challenges your views — even just to understand it;
- Diversify your information sources and platforms;
- Use impersonalized browsing periodically (e.g., incognito mode, with no logged-in accounts or saved data);
- Remember: algorithms optimize for attention, not truth.

B. Conversational bots for fraud or fake recruitment

„Not every natural conversation is human. Some are trained to deceive you.”

Description:

In this scenario, a seemingly legitimate AI-powered chatbot initiates a conversation with a user under a credible pretext — a job offers, financial assistance, professional mentorship, or personal coaching. The dialogue feels authentic, coherent, and empathetic — just what you’d expect from an HR specialist, a trusted colleague, or a reliable partner.

As conversation progresses, the user is gradually encouraged to provide personal information, documents, account access, or to take risky actions — all under the illusion of a sincere and professional relationship. Because AI simulates empathy and trust, the victim doesn’t realize they are interacting with a conversational manipulation system, not a human being.

AI techniques involved:

- Large Language Models (LLMs) – for natural, adaptive, and persuasive dialogue (e.g., ChatGPT, Claude, Gemini);
- Behavioral profiling – real-time analysis of the user’s language to personalize the message;
- Voice cloning AI (in automated calls) – mimicking the voice of a familiar person;
- Human-like interface simulation – avatars, names, logos, and credible automated responses.

Risks / Impact:

- Identity theft and sensitive data exposure – CVs, national ID numbers, addresses, personal documents, medical history;
- Financial fraud – fake recruitment fees, opening of fraudulent bank accounts;

- Corporate access breaches – if the victim provides login credentials during an “onboarding” simulation;
- Emotional manipulation – in some cases, the dialogue becomes emotionally charged and builds deep trust, especially with vulnerable users.

Where it happens:

- Professional networks (e.g., LinkedIn, job boards, career platforms);
- Direct messaging (e.g., emails, chats on social media, WhatsApp, Telegram);
- Chat interfaces on fake recruitment or company websites;
- During economic instability – when job promises carry greater emotional and financial weight.

Real-world example:

In 2023–2024, dozens of victims across Eastern Europe were contacted via Telegram by fake “IT recruiters” offering remote jobs. After a friendly and convincing chat, users were redirected to fake websites imitating known platforms and asked to upload personal documents. Behind it was a fully automated AI-driven conversation system, with no human intervention — a fact later reported in the media¹.

Warning signs for users:

- The recruiter avoids direct validation questions (e.g., “Who’s your supervisor?”, “Where can I call you?”);
- The language is polished but lacks concrete details about the role, company, or process;
- Refusals are met with empathetic insistence (e.g., “I totally understand, but...”, “This is a rare opportunity...”);
- You are quickly asked for documents or data without any official procedure.

How to protect yourself:

- Never send personal documents without verifying the recruiter’s real identity;
- Search for independent information about the company, job, and contact person;
- Look for generic phrases, inconsistencies, or subtle pressure in the messages;
- Avoid sensitive conversations with entities that refuse validation through multiple official channels (e.g., verified email, voice call, company website);
- If it seems too easy or too quick to be true — it’s probably an AI-driven scam.

C. Generation of hyper realistic fake media (Deepfakes)

„A picture is worth a thousand words. But what happens when that picture is a perfectly crafted lie?”

Description:

In this scenario, attackers use visual AI tools to generate fully fabricated yet highly realistic video or audio content, depicting real individuals (e.g., politicians, influencers, religious leaders, journalists, coworkers, etc.) in situations or statements they never actually made.

¹ EuroNews - Fake job offers :<https://www.euronews.com/next/2023/10/23/behind-the-global-scam-worth-an-estimated-100m-targeting-whatsapp-users-with-fake-job-offe>

Bitdefender - Beware of employment scams

<https://www.bitdefender.com/en-us/blog/hotforsecurity/8-telegram-scams-how-not-to-get-scammed>

The content is released at strategic moments — such as before an election, during a social crisis, or to discredit or support someone. Even if later debunked, the emotional impact often precedes rational verification, leaving lasting damage.

AI techniques involved:

- Deepfake video generation (e.g., face-swapping, lip-sync AI) – realistic facial and lip movement synchronization;
- Voice cloning – perfect imitation of a real person’s voice;
- Image/video generation tools (e.g., D-ID, Synthesia, DeepFaceLab);
- Text-to-video systems – turning written content into realistic-looking speeches or statements by synthetic presenters.

Risks / Impact:

- Mass disinformation – the public believes a false statement made by a fabricated “authority figure”;
- Blackmail and reputational damage – false videos used to intimidate or discredit individuals;
- Social panic – through fake declarations of war, pandemics, attacks, or terrorist acts;
- Erosion of trust in visual evidence – in the long run, people begin to distrust even legitimate footage (“everything can be faked”).

Where it happens:

- Political and electoral campaigns;
- Diplomatic tensions, military conflicts, social unrest;
- Personal or professional smear campaigns (e.g., against businesses, influencers, media);
- Fast-distribution platforms (e.g., TikTok, WhatsApp, Telegram, Facebook).

Real-world example:

As reported in various media outlets, a deepfake video circulated in 2022 allegedly showing the president of a country “announcing his surrender and stepping down.” The video was well-crafted and disseminated via partisan channels to demoralize the public. Though quickly debunked, millions had already seen and shared it².

Warning signs for users:

- Slightly unnatural facial movements, eye direction, or voice tone;
- High video quality from an obscure or unknown source;
- Shocking declarations not supported by official news;
- Exclusive circulation in closed groups or biased channels;
- Lack of original source or verifiable context.

How to protect yourself:

- Avoid sharing sensational materials without cross-checking from multiple sources;
- Use visual verification tools such as InVID, Deepware, Sensity;
- Compare the message with official transcripts, other video versions, or credible news outlets;

² France24 - Debunking a deepfake video of Zelensky telling Ukrainians to surrender
<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/>

- Pay attention to timing – if the release is too perfectly timed to provoke disruption, it may be synthetic;
- Educate your network – remind others that visual realism no longer guarantees authenticity.

D. AI-powered personalized influence messaging (Microtargeting)

„When AI knows exactly what to say, how, and when — so you believe you made the choice yourself.”

Description:

In this scenario, attackers or campaign operators use AI to craft highly personalized influence messages, precisely targeted at specific individuals or demographic groups. These messages are not just persuasive — they are engineered to trigger a specific emotional or behavioral reaction, whether it's a vote, a purchase, a political opinion, or a real-world action.

The message can take the form of a post, ad, article, video, or even a one-on-one conversation, delivered at the optimal time and in the right emotional context — all designed to maximize its manipulative effect.

AI techniques involved:

- Psychographic microtargeting – identifying the user’s cognitive style, values, and emotional vulnerabilities;
- Predictive neural networks – to forecast the most probable response to a given message;
- Adaptive content generation (e.g., personalized text, voice, video, imagery);
- Algorithmic delivery systems – adjusting the timing and frequency of message delivery in real time based on user behavior.

Risks / Impact:

- Manipulation of seemingly “free” choices, subtly steered by personalized stimuli;
- Electoral interference – voters are influenced differently depending on their emotional profiles;
- Exploitation of personal vulnerabilities – e.g., depression triggers “rescue offers,” fear triggers aggressive propaganda;
- Silent behavioral shaping – people are influenced without knowing it, leading to mass psychological alignment.

Where it happens:

- Political and ideological campaigns;
- Aggressive commercial advertising (including “miracle” products);
- Mass manipulation on social media;
- Targeted attacks on vulnerable demographics (e.g., elderly, youth, parents, emotionally distressed users).

Real-world example:

In the context of the 2016 campaigns and the Cambridge Analytica case, it was revealed that personal Facebook data was used to build psychographic profiles. An anxious voter received chaos-themed ads, a conservative one saw messages about lost values, while an undecided

voter was shown content about economic frustration. Each message was unique, but all converged toward the same voting behavior³.

Warning signs for users:

- You receive posts or ads that perfectly echo your own thoughts;
- You feel like “everyone agrees” with your opinion;
- You’re drawn to causes or ideas that were suggested to you when you were emotionally vulnerable;
- Others don’t seem to see the same content — the message is custom-tailored for you.

How to protect yourself:

- Don’t assume others see the same content — compare with independent sources;
- Avoid forming strong opinions based solely on ads, personalized feeds, or “coincidental” messages;
- Limit your digital footprint – don’t share too much personal data or take online personality tests;
- Use anti-tracking extensions, private browsers, and filters to reduce algorithmic targeting;
- Always ask yourself: “Why is this being shown to me — and why now, in this form?”.

E. Emotionally triggered manipulation (AI-driven exploitation of negative emotions)

„Anger, fear, and anxiety aren’t just reactions — they’re also tools.”

Description:

This scenario explores the intentional use of negative emotions (e.g., fear, anger, panic, shame, or moral outrage) as tools for algorithmic manipulation. Affective AI systems — capable of detecting users’ emotional states via behavioral analysis, facial expressions, voice tone, or digital interaction history — can be used to trigger and sustain such emotions in order to:

- Increase engagement,
- Influence rapid or impulsive decisions,
- Push users toward predefined actions (e.g., vote, donate, protest, purchase).

AI doesn’t generate emotions out of thin air — instead, it feeds them, reinforcing emotional states with matching content: alarming posts, negative news, aggressive videos, or morally triggering messages.

AI techniques involved:

- Affective computing / Emotion AI – real-time detection of emotional state;
- Emotionally adaptive feeds – delivering content that intensifies a user’s dominant emotion;
- Predictive behavioral modeling – identifying vulnerable moments (e.g., late at night, after failure, during crises);
- Emotion-based content targeting – rage bait, fear appeals, shame triggers.

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 Guardian - 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>

Risks / Impact:

- Manipulation of decisions under emotional pressure – impulsive shopping, reactive behavior, uncritical support of causes;
- Psychological destabilization – prolonged exposure to negative content leads to anxiety, depression, and paranoid thinking;
- Polarization and collective hatred – emotional exploitation fuels radicalization and social fragmentation;
- Increased susceptibility to scams and ideological manipulation – strong emotions reduce cognitive vigilance.

Where it happens:

- Political campaigns, health crises, social or environmental emergencies;
- Public scandals, disasters, terrorist attacks;
- Aggressive “fear-based” marketing;
- “Rage farming” campaigns on social platforms.

Real-world example:

During the COVID-19 pandemic, millions of users were exposed to AI-curated alarmist content (e.g., “the vaccine will kill you,” “they’re hiding the truth”) based on their interaction history. AI systems learned that fear drives longer watch times, more clicks, and mass sharing. The result: widespread panic, mistrust in authorities, and social division — as confirmed by multiple journalistic investigations⁴.

Warning signs for users:

- We consistently feel intense negative emotions after digital engagement (e.g., anger, fear, shame, outrage);
- We react impulsively, without analyzing the content logically;
- Our feed is dominated by dramatic, catastrophic, or emotionally loaded posts;
- The content reinforces our existing anxiety or distrust without offering nuance.

How to protect yourself:

- Limit emotional exposure during times of personal vulnerability;
- Train yourself to recognize algorithmic emotional traps: clickbait headlines, overly dramatic videos, urgency-based posts;
- Pause before reacting – avoid sharing, commenting, or acting while in a heightened emotional state;
- Fact-check with independent sources, especially if your emotional reaction is strong;
- Practice emotional literacy and critical thinking – if you feel too angry, someone likely got what they wanted.

⁴ The Guardian - ‘Alarming’: convincing AI vaccine and vaping disinformation generated by Australian researchers

<https://www.theguardian.com/australia-news/2023/nov/14/alarmed-convincing-ai-vaccine-and-vaping-disinformation-generated-by-australian-researchers>

The Trust & Safety Foundation - 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. Automated AI-powered spear phishing attacks

„You no longer need a skilled hacker – AI can launch mass-personalized attacks with surgical precision.”

Description:

In this scenario, attackers use artificial intelligence to automate spear phishing campaigns — targeted deception attempts crafted for specific individuals or small groups. Unlike traditional phishing, which relies on generic messages, AI-generated spear phishing is:

- Specific,
- Personalized,
- Highly convincing,
- Context-aware.

The AI system analyzes the target’s online presence (e.g., social media, articles, CVs, public interactions), then generates perfectly written messages and may even simulate real-time conversations to gain access, steal data, or request fraudulent transfers.

AI techniques involved:

- Large Language Models (LLMs) – generate context-tailored emails or chat messages (e.g., ChatGPT, Claude);
- OSINT-Based Profiling – AI scans public data about the victim (e.g., employer, colleagues, habits, interests);
- Identity Simulation – mimics the writing style of a colleague, client, or known contact;
- Voice Cloning / Audio Deepfakes – in some cases, AI clones a superior’s voice to deliver fake instructions by phone.

Risks / Impact:

- Identity or credential theft – victims unknowingly provide usernames, passwords, OTPs, or sign documents;
- Internal network compromise – attackers gain access to IT infrastructure through social engineering;
- Financial fraud – false wire transfers, payments to fraudulent accounts;
- Reputation damage or blackmail – stolen information used for coercion or professional sabotage.

Where it happens:

- Companies (e.g., HR, finance, IT, or C-level employees);
- Journalists, activists, political figures;
- Administrators with privileged access to systems;
- Geopolitical operations, corporate espionage, APT-level attacks.

Real-world example:

In 2023, cybersecurity researchers demonstrated that an AI system could generate a fully personalized spear phishing email — seemingly sent by a company’s CEO — in under 60 seconds. It matched the executive’s writing style and referred to real internal projects (sourced via public information). In testing, the click-through rate surpassed 70%, according to published reports⁵.

⁵ Since Direct – Spear phishing attack
<https://www.sciencedirect.com/topics/computer-science/spear-phishing-attack>

Warning signs for users:

- Receiving an unusual but well-written message from a known contact, containing a link, file, or urgent request;
- The message appears perfectly timed, referencing a recent project or using familiar phrasing;
- The sender pressures for immediate action: “urgent,” “just today,” “execute immediately”;
- Any hesitation is met with pseudo-empathic insistence: “I know you're busy, but please...”

How to protect yourself:

- Enable multi-factor authentication (MFA) for all critical accounts;
- Always verify suspicious requests through alternate channels (e.g., phone call, internal chat);
- Don't click links or open attachments without checking the full sender address and message context;
- Use anti-phishing filters, AI-based detection tools, and endpoint protection systems;
- Be aware: a perfectly written message is no longer proof of legitimacy — in fact, it might signal a highly advanced AI-crafted attack.

G. Simulated public consensus via AI bot networks

„When thousands of seemingly real voices say the same thing, you start to think you're the one who's wrong.”

Description:

In this scenario, attackers or manipulative actors use AI-controlled bot networks (social bots) to create the illusion of widespread social consensus. These bots simulate real users, complete with credible profiles, AI-generated images, activity histories, and persuasive posts.

The goal is to artificially amplify an idea, cause, outrage, or ideological narrative to the point where the public perceives it as:

- Mainstream,
- Reasonable,
- Inevitable.

This artificial social pressure has significant psychological effects: when “everyone” seems to support something, it becomes harder to question it—or even to hold a different opinion.

AI techniques involved:

- Fake identity generation (GANs) – hyperrealistic profile photos, entirely synthetic;
- LLMs – generation of posts, comments, replies, and messages that sound human;
- AI-driven orchestration – managing the simultaneous behavior of thousands or millions of accounts (posting, sharing, attacking, supporting);
- Conversational manipulation – emotionally and logically tailored replies that simulate debates between “diverse people”.

Risks / Impact:

- Fabricated trust in products, political messages, conspiracy theories, or smear campaigns;
- Silencing of real voices – overwhelming or discouraging dissent through volume and aggression;
- Social pressure and conformity – users begin self-censoring or adjusting their beliefs to fit the “majority”;
- Distortion of truth and genuine debate – online discussions become manufactured echo chambers.

Common contexts:

- Elections, referendums, political crises;
- Internal or foreign propaganda campaigns;
- Promotion of conspiracy theories, pseudoscience, or “miracle” products;
- Anti-Western, anti-EU, anti-NATO, or anti-democratic disinformation efforts.

Real-world example:

In 2020, social media platforms across multiple countries uncovered thousands of coordinated accounts spreading anti-vaccine and anti-lockdown messages. These fake profiles posed as concerned citizens, doctors, parents, or veterans, using AI-generated photos and false activity logs. A repeated slogan: “The people have awakened”⁶.

Warning signs for users:

- Many identical or very similar comments posted at the same time;
- Recently created profiles with no authentic activity or private interaction settings;
- Accounts focused obsessively on one topic, without variation;
- Rapid, coordinated replies attacking any differing viewpoint;
- A general sense that “everyone agrees”—with no nuance, critique, or real debate.

Protection measures:

- Check suspicious profiles (e.g., reverse image search, engagement history, robotic tone);
- Don’t let volume equal truth—ask: “Who are these people? Do they really exist?”
- Be wary of emotionally charged “crowd outrage” moments—ask: “Why now?”
- Don't change your beliefs just because it “seems” like the majority agrees—look for real arguments, not just noise.

H. Artificial public personalities for influence and manipulation

„Who's influencing you? A real person—or an AI entity with a hidden agenda?”

Description:

In this scenario, AI is used to create entirely fake public figures—influencers, experts, activists, or “credible voices”—controlled by an operator or organization. These personas are equipped with:

- AI-generated visuals (e.g., hyper realistic photos, animated avatars),

⁶ European Commission – Fighting disinformation

https://commission.europa.eu/strategy-and-policy/coronavirus-response/fighting-disinformation_en

Euro-Atlantic Resilience Centre - Barometer of societal resilience to disinformation

<https://e-arc.ro/wp-content/uploads/2022/05/Barometrul-rezilientei-societale-2022.pdf>

- Convincing biographies,
- Professionally written content (posts, videos, articles),
- Automated engagement with audiences.

The goal is to gain trust, build an audience, and gradually inject manipulative, ideological, or commercial messages into public discourse—with zero accountability, since the person isn't real and has nothing to lose.

AI techniques involved:

- Image generation (e.g., GANs, StyleGAN, Midjourney) – portraits, lifestyle shots, profile images;
- LLMs (e.g., ChatGPT, Claude, Mistral) – to generate posts, comments, articles, and personalized replies;
- Voice synthesis and video avatars (e.g., Synthesia, D-ID) – to produce realistic “talking head” videos;
- Bot networks – secondary accounts that amplify and validate the persona’s content.

Risks / Impact:

- Strategic opinion manipulation – personas gain trust and slowly distort public perception on sensitive topics;
- Fake “thought leaders” with no accountability or real identity;
- Imitation of authority professions (e.g., doctors, lawyers, journalists, humanitarians);
- Covert geopolitical influence – seemingly neutral figures pushing hostile agendas;
- Interference in civic, educational, religious, medical, or political spaces.

Common contexts:

- Social media platforms, private groups, video/streaming channels;
- Disinformation campaigns or ideological rebranding;
- Promotion of controversial products, pseudoscience, conspiracies, or political movements;
- Creation of fully AI-controlled “influencer” networks.

Real-world example:

In 2022, a network of “career women” influencers emerged on Instagram, promoting Western values in parts of the Middle East. Investigations revealed that all the accounts were AI-generated personas run by a government agency. Every post, reply, and image was synthetically produced, as covered in several in-depth reports⁷.

Warning signs for users:

- No evidence of the person outside the platform;
- “Too perfect” photos with vague or unverifiable backgrounds;
- Inconsistent or unverifiable biographical details;
- Excessively neutral tone, with no emotional variance;
- Hyperactive engagement (daily posting, 24/7 replies, instant comments).

⁷ PC Tablet - Embrace the Digital Wave: The Rise of AI Influencers
<https://pc-tablet.com/embrace-the-digital-wave-the-rise-of-ai-influencers/>
 You Dream AI – 10 examples of AI influencers on Instagram (the future is here)
<https://yourdreamai.com/ai-influencer-examples-on-instagram/>

Protection measures:

- Cross-check their existence through external sources (press, real-world events, authentic interviews);
- Be skeptical of “new influencers” who rise too fast with repetitive messaging;
- Don’t confuse social credibility (likes, comments) with authenticity;
- Be cautious when a “balanced” persona suddenly begins endorsing extreme, partisan, or toxic narratives—even if they once seemed trustworthy.

I. Orchestrated campaigns via mobile apps with embedded AI

„The app looks harmless. But behind the scenes, AI is orchestrating an invisible agenda.”

Description:

In this scenario, a mobile app that appears to be benign—such as one offering news, entertainment, education, spirituality, community engagement, or even health tracking—integrates covert AI mechanisms designed to manipulate information.

The app becomes a malicious platform through which distorted, false, or ideologically charged content is delivered with the intent to:

- Influence opinions,
- Steer emotions,
- Mobilize users into collective actions,
- Gradually radicalize specific groups.

Because the app is perceived as “trustworthy” (e.g. downloaded from official stores, well-rated, possibly backed by obscure but seemingly legitimate sponsors), the user suspects nothing.

AI techniques involved:

- LLMs – automatic content generation based on the user's profile and behavioral patterns;
- AI-controlled personalized feeds – real-time content adaptation according to user reactions;
- Emotion AI – detecting the user’s mood and adjusting messaging accordingly;
- Gamification with manipulative mechanics – rewards, points, or emotional incentives for radical or compliant behavior.

Risks / Impact:

- Disinformation delivered as “trusted content” – users skip verification, assuming the app is safe;
- Mass mobilization on false or inflammatory topics – protests, coordinated actions, mass reactions;
- Gradual ideological radicalization – from “soft” content to extreme beliefs via repetitive, adaptive exposure;
- Mass data harvesting for profiling – behavioral, social, political, or religious profiling without explicit consent;
- Creation of closed, self-reinforcing communities – resistant to alternative perspectives or factual correction.

Common contexts:

- Apps promoted as “alternatives” to mainstream media (“the truth others hide”);
- Apps targeting parents, alternative education, spirituality, or natural health;
- New messaging platforms claiming “absolute freedom of speech”;

- Campaigns indirectly sponsored by political actors or opaque influence groups.

Real-world example:

In multiple countries, mobile apps claiming to deliver “uncensored news” were revealed to be controlled by partisan propaganda networks. Users were gradually exposed to false narratives about global elites, public health conspiracies, or calls to rebellion—disguised behind a clean interface and professional tone, as confirmed by media reports⁸.

Warning signs for users:

- The app offers “exclusive truths” or promises to “wake you up”;
- Vague or absent sources, repeated references to “hidden systems that lie to us”;
- Increase in negative emotions after use (e.g. frustration, fear, distrust of all institutions);
- Recommendations lead to closed groups, radical forums, or “urgent action” calls;
- Frequent push notifications about crises, betrayals, conspiracies, etc.

Protection measures:

- Check the developer and origin of the app—who controls it, and what its goals are;
- Confirm whether the information provided is supported by independent sources;
- Pay attention to your emotional response to the app—frequent negativity may be a manipulation signal;
- Uninstall apps that offer only one-sided narratives and promote absolute distrust in everything else;
- Learn to spot fear-, hate-, or superiority-based narratives masquerading as truth.

J. AI-Based influence in education – platforms, “mentors,” and distorted learning resources

„Not all lessons come from textbooks—and not all teachers are human or unbiased.”

Description:

In this scenario, AI is misused to negatively influence the education of young people or general audiences via distorted, biased, or entirely false content delivered through:

- E-learning platforms,
- Educational apps,
- Mentor-like AI chatbots,
- AI-generated educational videos,
- “Alternative courses” marketed as more “authentic” than official curricula.

While these tools appear innovative or helpful, they are intentionally designed to push manipulative narratives, unfounded theories, or ideological content disguised as hidden truths.

AI techniques involved:

- LLMs – automated generation of answers, explanations, and lessons tailored to students’ queries;
- Text-to-video + AI avatars – video lessons delivered by lifelike but entirely fake “teachers”;
- Adaptive learning systems – tailoring content based on the student’s cognitive style and emotional state;
- Educational microtargeting – delivering different resources based on ideological profiling.

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

Risks / Impact:

- Misinforming young minds – repeated exposure to falsified or ideological content shapes flawed reasoning;
- Eroding trust in formal education – replaced by unregulated “alternative systems”;
- Spreading conspiracy theories and pseudoscience under the guise of education;
- Early-stage ideological polarization – children trained to reject certain values, theories, or scientific institutions;
- Shaping generations conditioned for digital obedience, not critical thinking.

Common contexts:

- Unaccredited e-learning platforms that become popular with youth;
- General knowledge video channels with hidden agendas;
- “Personal development” apps that slide into dogma or radical activism;
- Homework chatbots providing biased, incorrect, or speculative answers.

Real-world example:

- In 2023, UNESCO issued a warning about how AI poses a risk to collective memory of the Holocaust. They documented how some AI models—search engines, conversational systems, and generative tools—returned inaccurate or revisionist results when users searched for Holocaust information, a concern highlighted on their official website⁹.

Warning signs for users:

- The app/platform has an “alternative” tone but completely rejects accredited academic systems;
- Lessons frequently contain phrases like “what no one wants you to know,” “hidden truth,” or “schoolbooks lie”;
- AI answers are delivered with authority, but without sources;
- The digital “teacher” repeatedly promotes a single ideological, anti-scientific, or conspiratorial stance;
- Feedback discourages critical thinking, pushing instead for uncritical acceptance of a narrative.

Protection measures:

- Use transparent, accredited educational platforms with verifiable content;
- Ask AI for sources and verify them independently;
- Don’t rely on a single educational tool—compare answers and cross-check materials;
- Encourage debate, questions, and constructive doubt—don’t passively accept “delivered lessons”;
- Train students in media literacy and AI literacy to spot manipulative content.

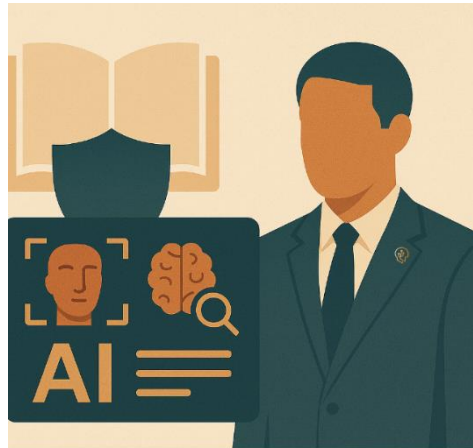
⁹ UNESCO - AI and the Holocaust: rewriting history? The impact of artificial intelligence on understanding the Holocaust

<https://www.unesco.org/en/articles/ai-and-holocaust-rewriting-history-impact-artificial-intelligence-understanding-holocaust>

5 PREVENTION METHODS

Prevention in the AI era is no longer just about installing antivirus software or avoiding shady websites. It now involves cultivating healthy digital habits, developing critical thinking, and understanding the algorithmic systems that shape our daily online experiences. We are not only facing a technological issue—but a cognitive and social one.

This chapter offers a selection of practical, non-exhaustive, yet relevant measures for individual users, educators, institutions, and technology developers.



The goal is not only passive protection, but the creation of a culture of digital vigilance—where every user becomes active, critical, and aware of the invisible mechanisms that can shape their perception and behavior.

5.1 For individual users

„You’re not powerless against algorithmic manipulation – but you must learn to recognize it and respond wisely.”

Artificial intelligence can manipulate subtly, persuasively, and invisibly. But the general public has concrete, effective methods to protect their cognitive autonomy and trust in reality. This section presents a set of simple but crucial practices for recognizing, resisting, and countering AI’s abusive influence on perception and behavior.



A. Train your digital critical thinking

Your first and most important filter against manipulation is your own discernment.

- Always ask yourself:
 - Who wants me to believe this?
 - Who benefits if I react emotionally or impulsively?
 - Why is this information appearing right now?
- Don't rely on first impressions – AI is trained to serve content that immediately “hooks” you: sensational headlines, personalized messages, shocking visuals. Learn to take a step back and rethink your reaction.
- Train your reflex to analyze, not just react. Critical thinking is a form of digital self-defense.

B. Check the source and context of the content

Information without a source, context, or identifiable author can be more dangerous than an openly declared lie.

- Avoid emotional, impulsive reactions. If something makes you instantly angry, anxious, or “hit by the truth,” that’s a red flag—it might be manipulative.
- Check:
 - Who published this content?
 - Is the author real, known, verifiable?
 - When and in what context did this message appear?
 - How is it being shared and by whom?
- Use external, neutral sources for confirmation. Don't trust only what “shows up”—actively seek alternative perspectives.

C. Recognize algorithmic manipulation

If you're constantly seeing the same type of content, you may not be informed—you may be trapped in a digital pattern.

- If your feed feels too “uniform” or repetitive, ask yourself: *Where are the opposing views? Why don't I see them?*
- Actively seek contrast:
 - Explore ideologically opposing sources
 - Compare headlines
 - Talk to people with different perspectives
- Diversify your sources:
 - Don't rely on a single platform
 - Use different search engines, independent outlets, international sources.
- Don't let AI decide what you see—take back control of your own information intake.

D. Use AI / deepfake detection tools

Appearances can be machine-made. Be more vigilant than a pixel.

- When encountering suspicious videos, audio, or images, use specialized tools:
 - Deepware Scanner – detects deepfake video/audio
 - Hive AI – automated visual and audio analysis
 - Sensity AI – enterprise-grade solutions for visual manipulation detection

- Microsoft Video Authenticator – checks video authenticity
- Look for telltale inconsistencies:
 - Rigid or unnatural facial expressions
 - Flat or overly robotic voices
 - Poor lip-sync
 - Unrealistic gestures, generic or repeated backgrounds.
- Use reverse image search (e.g., Google Images, Yandex) not only to verify if an image has been used in another context, but also to detect manipulation and misinformation.

5.2 For organizations

„In an era where a single fake video can destroy your reputation and internal decisions can be influenced by a chatbot, organizations must defend themselves intelligently and proactively.”

Organizations—such as companies, public institutions, NGOs, educational structures, or security bodies—are priority targets in AI-based manipulation strategies. Targeted disinformation, conversational fraud, and public image sabotage can cause severe financial, operational, trust, and reputational damage. Effective prevention requires a combination of systemic, technical, and cultural measures..

A. AI manipulation awareness & defense training

Training staff is the **first line of defense** against cognitive attacks and sophisticated social engineering tactics.

- Internal training programs for employees, PR, HR, IT, legal, and executive management on:
 - Identifying algorithmic manipulation;
 - Deepfake risks (video, audio, text);
 - Recognizing AI-based phishing and conversational fraud.
- Cognitive attack simulations:
 - Scenario testing with fake videos of “executives”;
 - AI-generated spear phishing emails;
 - Bot-driven recruitment scams or fake financial request simulations.
- Internal rapid response guides:
 - What to do if a deepfake of the CEO surfaces?
 - How to verify urgent requests sent via “credible” channels?
 - What to communicate publicly and how to preserve trust?

B. Multi-channel validation policies

In a volatile digital environment, critical decisions should never rely on a single communication channel.

- Any financial, contractual, or strategic decision must be:
 - Double-verified through two or more independent channels (e.g., email + phone call + in-person confirmation);
 - Cross-authenticated, especially if coming from unusual sources or outside working hours.

- Video calls and audio messages are no longer reliable proof, given how realistic facial and voice cloning has become.
- Internal procedures should be updated to ensure no single department has sole decision-making power in critical cases without multi-point verification.

C. Automated and manual reputation monitoring

Organizational reputation is a primary target in information warfare. A well-coordinated attack can damage trust in a matter of hours.

- Implement automated monitoring tools for mentions of brand names, key executives, products and services, especially on:
 - Social media;
 - Messaging apps (e.g., Telegram, WhatsApp groups);
 - Alternative sources (Dark Web, fringe forums);
 - Video platforms and fake-news outlets.
- Detect coordinated AI-based campaigns:
 - Simultaneous posts, artificial accounts, identical wording;
 - Deepfakes mimicking official statements;
 - Fabricated documents that appear “leaked”.
- Rapid response teams for reputation management:
 - transparent and direct information campaigns for the public, partners, and the press.

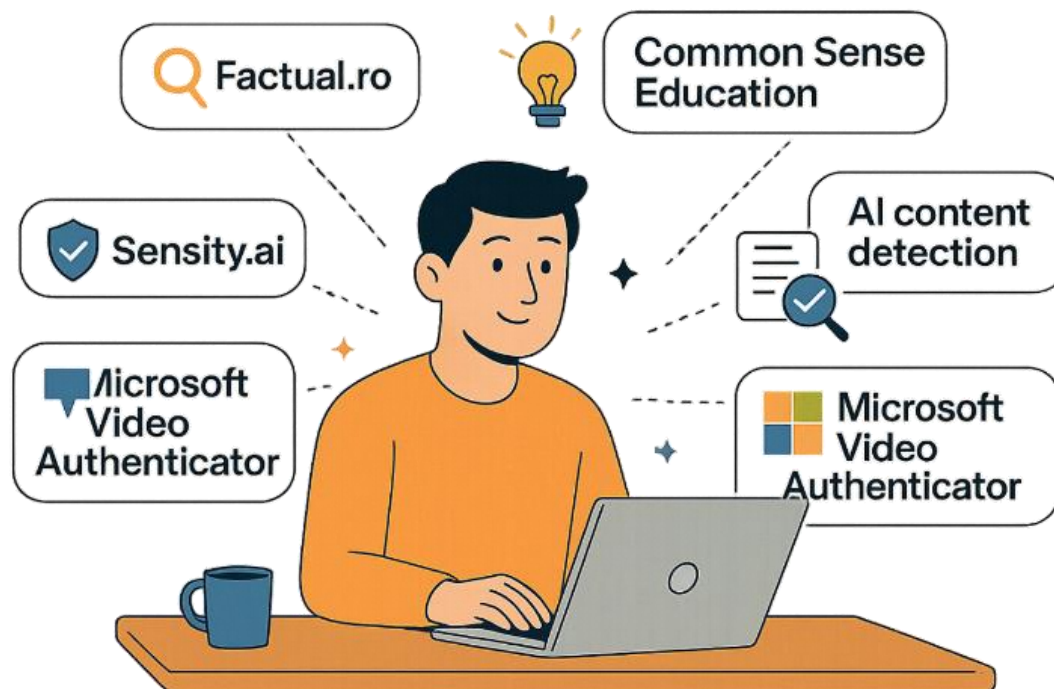
D. Collaborate with experts, fact-checkers & specialized organizations

Effective defense is collective by nature. No single entity can detect, analyze, and respond to today’s sophisticated AI manipulation alone.

- Establish partnerships with:
 - Investigative journalism and fact-checking teams;
 - Experts in cybersecurity, social psychology, and crisis communication;
 - AI detection platforms (e.g., Sensity, Deepware, Graphika);
 - NGOs monitoring the information space and disinformation trends.
- Gain access to early-warning networks, including national CERTs or OSINT groups, to respond quickly to deepfake campaigns or reputational attacks.
- Participate in collective initiatives for digital resilience: public education campaigns, online safety training programs, best practice frameworks for digital defense.

6 USEFUL RESOURCES AND ADDRESSES

„Access to accurate information and verification tools is the first line of defense against AI-based manipulation.”



In an information landscape increasingly dominated by AI-generated content, it is essential that the general public, educators, professionals, and organizations know and use verified resources and effective tools. Below are several useful platforms for fighting disinformation, promoting critical education, and detecting false content created with AI:

Fact-checking for Romania - <https://factual.ro>

A Romanian-language fact-checking platform dedicated to debunking false claims in the public space.

- Analyzes and classifies statements from politics, media, and social networks;
- Provides sources, context, and explanations for verdicts (e.g., true, false, partly true, etc.);
- Extremely useful for developing the habit of verifying information, especially in electoral and socially sensitive contexts.

AI-generated content detection (deepfake, fake visual media) - <https://sensity.ai>

A professional visual security platform specialized in detecting AI-generated manipulations.

- Detects deepfake videos, doctored images, voice cloning, and visual media fraud;
- Offers advanced solutions for organizations, media, public institutions, and corporations;
- Can also be used for educational purposes, to concretely demonstrate how visual manipulation work.

Experimental AI tools by Google - <https://ai.google/tools>

A collection of AI-based applications and experiments, open to the public.

- Enables understanding of AI mechanisms in an interactive and safe manner;
- Includes tools for text, image, sound generation, and automated translation;
- Useful for introductory AI courses, digital literacy, and critical analysis.

EU-focused disinformation monitoring - <https://www.euvsdisinfo.eu>

An initiative by the European External Action Service (EEAS), dedicated to exposing and countering disinformation campaigns.

- Offers a database with examples of false narratives, sources, and propagation channels;
- Analyzes thematically and geographically how disinformation affects EU member states;
- An important tool for journalists, educators, fact-checkers, and strategic communication professionals.

Educational resources for media literacy and critical thinking - <https://www.commonsense.org/education>

A non-profit platform offering free resources for educators, parents, and students, focused on developing critical thinking and digital responsibility.

- Includes structured lessons on fake news, media bias, social influence, and online responsibility;
- Tailored for different age groups, with videos, worksheets, and teacher guides;
- Can be integrated into curricular or extracurricular activities focused on media and AI education.

Other recommended tools (for quick use):

- InVID Plugin – a browser extension for video and image analysis;
- Deepware Scanner – checks authenticity of video/audio files;
- NewsGuard – automatically evaluates the credibility of news websites;
- WhoTargetsMe – visualizes and analyzes political ads targeted at users on social media.

Recommended educational resources

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>

Romanian Intelligence Service - National Cyberint Center

Intelligence

- <https://intelligence.sri.ro/>

Buletin Cyberint

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

Romanian National Cyber Security Directorate (DNCS)

Deepfake and Social Engineering

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

Deepfake detection

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

Romanian Police

Deepfake used by cybercriminals

- <https://sigurantaonline.ro/deepfake-utilizat-de-infracatorii-cibernetici-pentru-promovarea-unor-opportunitati-false-de-investitii-pe-retelele-sociale/>

Online fraud awareness quiz

- <https://quiz.sigurantaonline.ro/>

Cloud Security Alliance Romanian Chapter (CSA_RO – part of CSA)

AI Organizational Responsibilities: AI Tools and Applications

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

Dynamic Process Landscape: A Strategic Guide to Successful AI Implementation

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

AI Controls Matrix

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

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>

Cyber Security Cluster of Excellence

Cyber Security

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

7 PREPARING FOR THE ALREADY-PRESENT FUTURE

In this new digital ecosystem, risk no longer stems only from disinformation or external attacks, but also from constant exposure to personalized, emotional, and often manipulative content. That is why prevention is no longer just about technology, but about a conscious, daily-practiced digital hygiene.

To cope with this reality, the following five fundamental directions are fundamental:

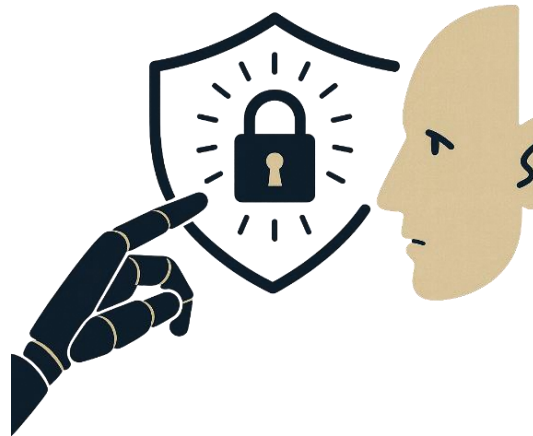
AI-era adapted digital education

Traditional online safety education must evolve into a new paradigm: perceptual digital literacy. This involves training users – from students to decision-makers – to recognize subtle

manipulation, identify AI-generated content, and understand how algorithms influence attention, emotions, and beliefs.

In schools and institutions, programs should include:

- Concepts about algorithmic personalization;
- Distinguishing between human and simulated interaction;
- Exercises in critically analyzing digital sources



Clear and updated regulations

Content-generating technologies evolve far faster than legislation. That's why it's vital to adopt clear legal frameworks that:

- Ban or regulate the use of manipulative AI-generated content (e.g., deepfakes in election campaigns);
- Require platforms to transparently label automatically generated content;
- Enforce accountability for AI developers regarding the negative impacts of their applications.

These regulations must protect both individual rights and democratic balance.

Algorithmic transparency and ethical audits

AI systems capable of influencing human behavior (e.g., social media platforms, search engines, chatbots) must be subject to independent audits. The public has the right:

- To know what personal data is being analyzed;
- To understand why certain types of content are shown;
- To opt for an algorithm-free feed.

Institutions should also support the development of mandatory ethical standards for AI, especially in education, health, justice, and politics.

Multidisciplinary collaboration

Perceptual manipulation is not just a technical issue. An integrated approach is needed, involving collaboration between:

- Cybersecurity and AI specialists;
- Psychologists and neurologists (to understand emotional responses);
- Educators and trainers (for critical dissemination of information);

- Lawyers and digital rights experts;
- Ethicists and sociologists (for social impact analysis).

Only through such cooperation can we understand AI's real effects on society and build effective protection mechanisms.

Accessible detection and verification tools

Just as every user has access to a search engine or browser, in the near future they should also have access to:

- A deepfake detection tool installed on their phone or laptop;
- A browser extension that flags AI-generated content;
- An app for quick verification of sources or content authenticity.

8 CONCLUSIONS

Artificial intelligence is no longer an emerging technology – it's an invisible force shaping more and more of what we think, feel, and choose. In a hyper-personalized digital ecosystem, where content is filtered, emotions are measured, and reactions are anticipated, the risk of subtle but systematic influence is a reality we face daily – often without realizing it.

Information manipulation no longer looks like it used to. It is not loud, obvious, or crude. It is finely calibrated, contextual, and personalized – an algorithm that knows what to say, when to say it, and in what tone, to provoke the desired response. And the source of these adjustments is often our own digital behavior: what we search, what captures our attention, what scares or comforts us.

This paper aimed to provide a clear overview of how AI can become a tool for perceptual shaping – through technology, psychology, and conversational design. More than a warning, it offers principles of digital hygiene and critical thinking, helping transform the passive user into a conscious actor of their own informational reality.

9 GLOSSARY

Technology and AI

- Artificial Intelligence (AI) – The simulation of human cognitive processes by computer systems capable of learning, reasoning, and making autonomous decisions.
- Artificial Neural Networks (deep learning) – Algorithmic architectures inspired by the human brain, used to recognize complex patterns in texts, images, or voices.
- Machine Learning – A branch of AI that enables systems to learn and evolve without being explicitly programmed.
- Large Language Models (LLMs) – Natural language processing models trained on vast datasets to generate and interpret text (e.g., ChatGPT, Gemini).
- Emotion AI / Affective Machine Learning – AI specialized in detecting and interpreting users' emotional states.
- NLP (Natural Language Processing) – Technologies that allow machines to understand and generate human language.
- GANs (Generative Adversarial Networks) – Networks capable of generating realistic-looking images, sounds, or videos.
- Face Swapping – A technique for replacing a person's face in a video or image with that of another.
- Voice Cloning – Artificial reproduction of a real person's voice using AI.
- Lip-syncing AI – Synchronizing lip movements in a video to match a generated or altered voice.
- Synthmedia / Synthetic content – Media content entirely generated by AI, without human input.
- Synthetic avatars / AI avatars – AI-generated animated graphical representations that can mimic real people.
- Text-to-image models – AI that generates images based on text descriptions.
- Motion capture AI – AI technologies that replicate body movements to realistically animate avatars.
- Tacotron / WaveNet – AI systems for voice synthesis with natural intonation and accent.
- Midjourney / DALL·E / Stable Diffusion / LLaMA / Claude / Gemini / ChatGPT / Mistral – Names of advanced AI models used for generating text, images, or simulated conversations..

Digital manipulation

- Perceptual manipulation – The invisible influence over how a person perceives reality, through personalized or simulated content.
- Algorithmic manipulation – Steering user behavior through automated selection of displayed information.
- Information bubble – A personalized digital space where the user only receives content that confirms their existing beliefs.
- Psychographic microtargeting – Delivery of emotionally personalized content based on a user's psychological profile.
- Automated spear phishing – Personalized phishing attacks using AI-generated deceptive messages that appear to come from known individuals.
- Advanced social engineering – Use of AI for complex psychological manipulation aimed at fraud, control, or influence.
- Predictive behavioral recommendation – Using AI to anticipate and shape user decisions.

- Content filtering – Automatic exclusion of alternative viewpoints to reinforce a specific perception.
- Information polarization – Dividing users into ideologically opposing groups through targeted content.
- Digital radicalization – The process by which AI fosters extreme beliefs through repeated exposure to radical content.
- Simulated social consensus – Artificial creation of the impression that an opinion is widely supported.
- Simulated empathy / Empathic chatbot – Chatbots that mimic human empathy to gain trust and influence.
- AI influencer / Artificial influencer – Social media accounts controlled by AI that simulate real people to generate influence..

Education and digital security

- Digital critical thinking – The ability to objectively analyze and evaluate digital content.
- Cyber education – Training on online risks and protection mechanisms.
- Fact-checking – The process of analyzing information to verify its accuracy.
- Algorithmic auditing – Systematic evaluation of how an algorithm works and influences users.
- Algorithmic transparency – The user’s right to know how data is processed and why certain content is shown.
- Verification reflex – The automatic reaction to validate information before believing or sharing it.
- Information hygiene – A set of practices to maintain healthy and balanced information consumption.
- Information self-defense – A set of skills and techniques through which users protect themselves from manipulation and disinformation.
- AI-generated content – Any material created automatically by artificial intelligence.
- AI ethics – The branch that analyzes the moral implications of developing and using artificial intelligence.

10 BIBLIOGRAPHY

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>

Euro-Atlantic Resilience Centre - Barometer of societal resilience to disinformation, <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 surrender, <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/cmcmds/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/alarmed-convincing-ai-vaccine-and-vaping-disinformation-generated-by-australian-researchers>

The Guardian - 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 Foundation - 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>