ANALYZE – DECIDE - ACT

Artificial Intelligence
& The Manipulation of Human Perception

The educational program
"Analyze – Decide – Act"

Mircea Constantin ȘCHEAU

Alexandru Ciprian ANGHELUȘ

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.

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

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.

Authors

The Manipulation of Human Perception has gone through different stages throughout history and in this sense, Artificial Intelligence is proving to be an effective and potent tool in the hands of increasingly active criminal groups. A correct analysis and a good source of information can be the ingredients for success in the legitimate effort to counter malicious actors, limit vulnerabilities and mitigate the negative impact and damage.

The work is balanced, well-structured and offers a realistic and clear perspective on the phenomenon. Even in the context of a rapid evolution of information technologies, the mechanisms of emotional alteration remain, however, the same. Cognitive filtering of fake content may be an answer to the avalanche of algorithms and conversational bots. Training in digital critical thinking is necessary to avoid predictive behavioral profiling and malicious shaping of user decisions.

The analyses are intended for the public, professionals and laypeople alike. The guide emphasizes the practical side of the issues and refers to sources provided by relevant organizations and institutions with responsibilities in the field. The joint effort led to the development of this study, one of the goals being the rapid increase in the level of awareness and, through this, the strengthening of the cybersecurity culture and resilience.

The Romanian National Cyber Security Directorate (DNSC)

Cybercrime is constantly evolving, amid the accelerated development of digital technologies and the expansion of internet use in everyday life. These transformations bring significant benefits but also increased risks for users.

Artificial Intelligence (AI) is having an increasingly pronounced impact on society, influencing the way we interact in the digital environment. Understanding the basic mechanisms of AI and how it can be used for fraudulent purposes is essential in preventing cybercrime. This guide, developed in support of prevention activities, is addressed to both the public and professionals, providing useful guidelines for the responsible and safe use of digital technologies.

The Institute for Crime Research and Prevention

TABLE OF CONTENTS

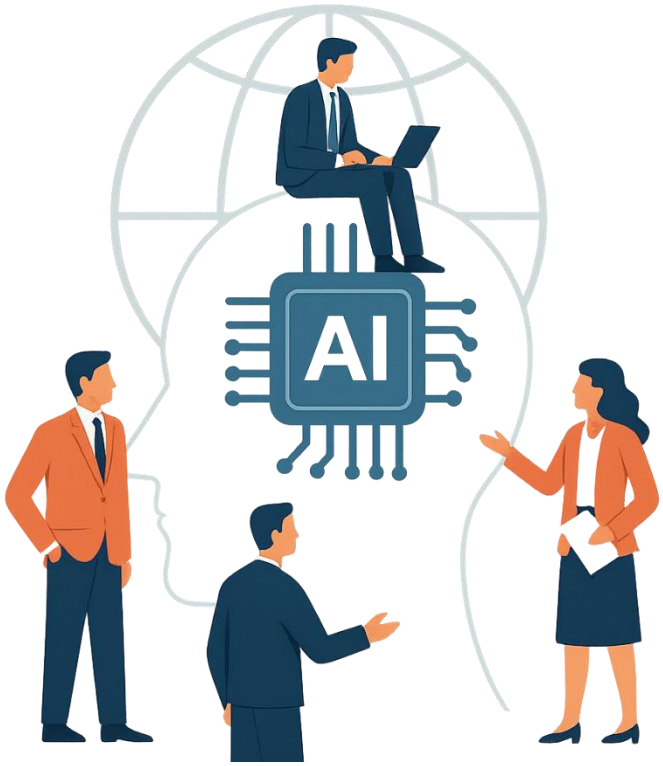
1. GENERAL CONCEPTS

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:

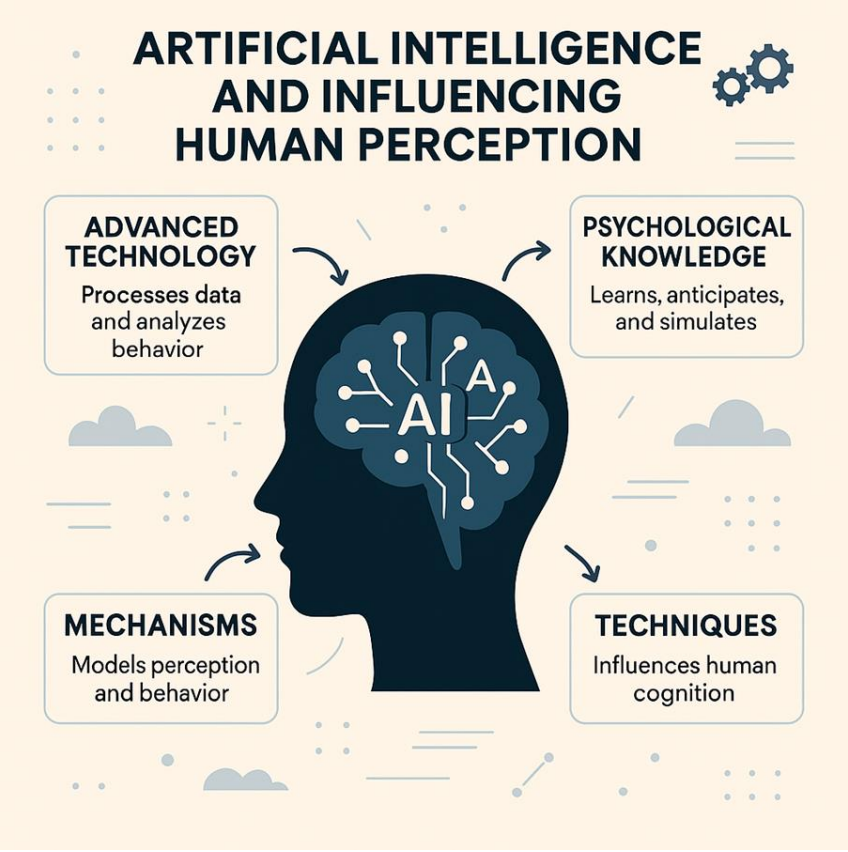
- What is a recommendation algorithm, and how does it influence what I see?
- How can a neural network “understand” my emotions?
- What happens to the data I provide — consciously or unconsciously?
- Why does some content feel like it was “made just for me”?

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 jurnalists).
- 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

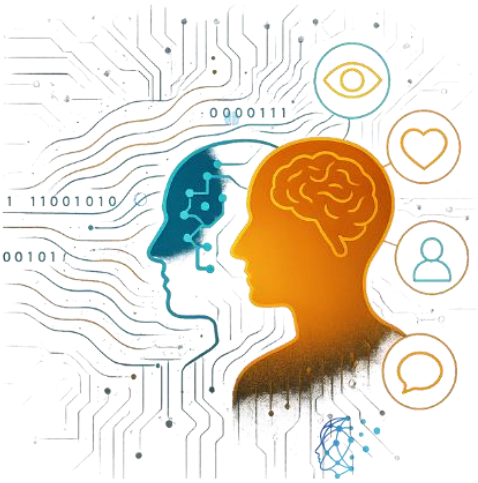
How can we protect ourselves?

- 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 TROUGH 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

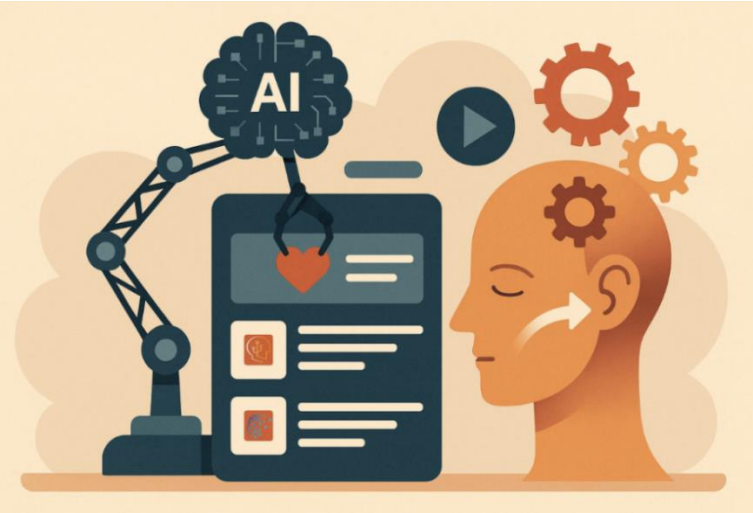
Common manifestations

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?

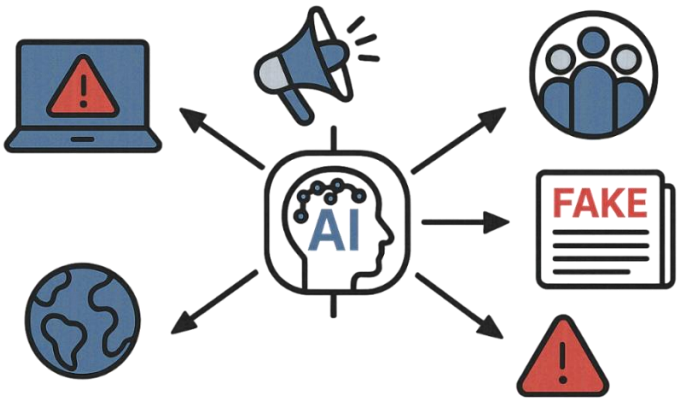
- Minimize behavioral data sharing: disable history tracking, restrict cookies, use secure/private browsers;
- Acknowledge that AI may know your patterns better than you do;
- Avoid making major decisions when tired, stressed, or emotionally charged.
- Ask yourself: "Is this truly what I want — or was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

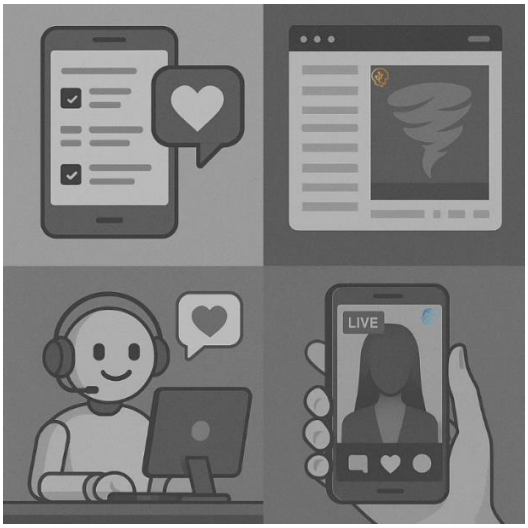
- Access to massive personal and behavioral data:
- AI can process real-time information about who you are, what you think, how you feel, and how you react — creating highly detailed personal profiles.
- Realistic content generation and simulation:
- AI can produce text, voices, images, and videos that 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 media[1].

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 it[2].

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

- 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 behavior[3].

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 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

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 investigations[4].

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/alarming-convincing-ai-vaccine-and-vaping-disinformation-generated-by-australian-researchers
The Trust & Safety Fundation - AI-Generated Disinformation Campaigns Surrounding COVID-19 in the DRC
https://www.trustandsafetyfoundation.org/blog/blog/ai-generated-disinformation-campaigns-surrounding-covid-19-in-the-drc

F. 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 reports[5].

[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”.

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

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 reports[7].

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 reports[8].

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 website[9].

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:

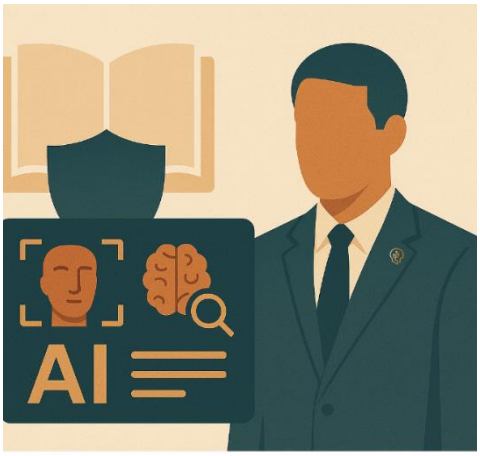
- Use transparent, accredited educational platforms with verifiable content;
- Ask AI for sources and verify them independently;
- Don't rely on a single educational tool—compare answers and cross-check materials;
- Encourage debate, questions, and constructive doubt—don't passively accept "delivered 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 anipulation

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 inconsistencie:
    - 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.

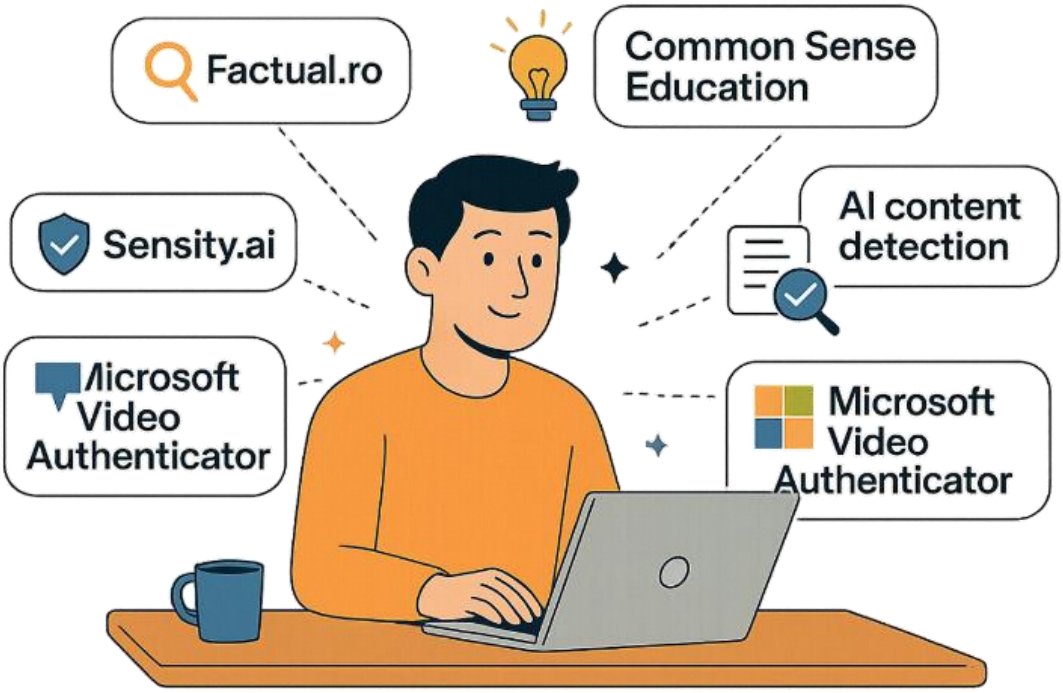
- Establish partnerships with:
  - Investigative journalism and fact-checking teams;
  - Experts in cybersecurity, social psychology, and crisis communication;
  - AI detection platforms (e.g., Sensity, Deepware, Graphika);
  - NGOs monitoring the information space and disinformation trends.
- Gain access 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 educatio.

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 (DNSC)

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-infractorii-cibernetici-pentru-promovarea-unor-oportunitati-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 surrende, https://www.france24.com/en/tv-shows/truth-or-fake/20220317-deepfake-video-of-zelensky-telling-ukrainians-to-surrender-debunked

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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