understanding europe

Beyond the Algorithm: Digital Resilience in the Age of AI



Schwarzkopf Foundation
YOUNG EUROPE

WORKSHOP

Beyond the Algorithm: Digital Resilience in the Age of AI

AIMS

› Understand **what artificial intelligence (AI)** is and where it shows up in our lives.

› Identify **AI-related issues**, their **impact** in young people's lives, and their ability to practice critical digital citizenship.

› Map frameworks and attitudes young people can use to build their sense of agency and self-efficacy towards **actional digital resilience.**

MATERIAL NEEDED

› Crash Course PowerPoint to be obtained via the Toolbox at www.understanding-europe.org

› Flipcharts.

› Markers and pens.

› Post-its.

› Scenario cards (ideally printed, alternatively digital).

DURATION

Short version: ~ 40 minutes.

Long version: ~ 60 minutes.

INTRODUCTION

This workshop explores digital resilience in an AI-driven world using hands-on and solution-oriented methods. The goal is to understand what AI is, where it shows up in our lives, and how to build digital resilience.

The workshop is grounded on examples of AI in young peoples' daily lives so the participants can learn the scope of this new technology. The workshop also challenges participants to reflect on AI-related issues, from deepfakes and misinformation to environmental impact and accountability. By using concrete fictional scenarios relevant to young people and highlighting tangible steps towards digital resilience, the focus is on promoting critical thinking and a sense of agency.

Participants are invited to identify the frameworks and attitudes that allow them to navigate this increasingly digital world with confidence. By the end, they should be able to recognise how they can contribute to a safer and more resilient digital space as critically active young citizens.

GENERAL OVERVIEW OF THE WORKSHOP

DURATION SHORT ~ 40 min	DURATION LONG ~ 60 min	MODULE	DESCRIPTION
10-15 min	20 min	1. What is AI?	Goal: Understand what AI is and where it shows up in our lives.
		1.1 “AI or not” - Defining AI	Different examples of possible AI applications are given to the participants. They place themselves along an imaginary line in the room depending on whether they think there is AI involved. At the end, AI is defined objectively.
		1.2 Other relevant terms	Through an interactive quiz, the trainer defines algorithm, machine learning, generative AI, and large language model (LLM) with the class.
--	15 min	2. Identifying AI-associated issues	Goal: Get an overview of AI-associated issues. The participants brainstorm issues related to AI and cluster them around categories.
20 min	20 min	3. Building digital agency in an AI world	Goal: Explore the impact of AI-related issues in young people’s lives and map steps they can take to build digital resilience.
		3.1 10 Steps to Digital Resilience	The trainer presents a brief intro to the topic, including frameworks and attitudes for digital resilience in an AI-driven world.
		3.2 Map of actionable digital resilience	Using this as a starting point, participants analyse a scenario and make their own map of actionable digital resilience by identifying where is the AI, its impact, and action points.
5 min	5 min	4. Committing to one small action	Goal: Moving from big picture back to their day-to-day life. In an informal round, participants commit to one small action to take in the immediate future to build their digital resilience.

MODULE 1:

1.1 AI OR NOT?

In the short version, 1.1 should be skipped. Trainers jump straight to the definitions in 1.2.

Duration: Approximately 10 minutes (long version)

Goal: Highlight different places where AI applications appear.

The trainer shows the participants 3-4 examples of applications which may be using AI. Participants can then place themselves along an imaginary line in the room depending on whether they think there is AI involved or not.

EXAMPLE	AI OR NOT?
ChatGPT (by OpenAI) Gemini (by Google), or Claude (by Anthropic), or another LLM	**This is AI.** Large language models (LLMs) are autonomous agents capable of reasoning, planning, using tools, and adapt based on what they learn from the interactions.
Spotify / Instagram recommendation algorithm	**This is AI.** Recommender systems learn from millions of users' listening behaviour to improve recommendations and constantly adapt their recommendations.
E-mail spam filter	**This is AI.** The system automatically identifies potential spam. It uses statistics, analysing the text for words that are more/less likely to be in spam vs. real emails.
GPS navigation system	**It depends.** Normal signal processing and geometry used to determine coordinates (i.e., how the GPS "knows" where we and our destination are) isn't AI. Giving good suggestions for navigation, especially when it adapts to variables like traffic conditions, CO_2 consumption, or learned past user preferences, requires AI features.
Brightness and colour adjustment features on photo editing apps	**It depends.** Editing an image can be done manually. However, automatic edition functions may use AI.
Phone calculator	**Not AI.** It's a fixed algorithm. It only calculates the formulas we type in.
Excel spreadsheet calculating formulas typed in by the user	**Not AI.** The outcome is determined by user-specified formulas.

During the activity, the trainer may write down keywords, such as the ones underlined above, on a flipchart or classroom board to help with the next step of the module.

NOTE FOR THE TRAINER

Choosing your examples:

- Choose examples that include more clear AI applications (e.g. social media algorithms) and mixed ones (e.g. GPS navigation).
- Choose examples that include more polarizing AI applications (e.g. ChatGPT) and less polarizing ones (e.g. spam filters).

Moderating the discussion:

- Highlight the fact that **AI is not a new concept**, though it did go through a boost in development, followed by a boost in public discourse.
- Underline the fact that, though AI is now heavily linked with LLMs (e.g. ChatGPT, Gemini, Claude), it involves many **more and diverse applications.**
- Point out that **not all AI applications are as polarizing as LLMs.**

OTHER NOTE

Things are changing fast, so it might be relevant to update the examples in this workshop from time to time.



In the short version, as you skipped 1.1, ask them to give you 2-3 examples of AI applications first.

MODULE 1:

1.2 DEFINING AI

Duration: Approximately 10-15 minutes

Goal: Give an objective definition of AI and other key concepts.

Now the trainer gives an objective definition of AI, relating it to the examples in 1.1.

Artificial intelligence (AI) is a field of study, like computer science or statistics; what falls within it (or not) is therefore constantly evolving.

- The previous examples are not "AIs", but AI models or systems.
- AI systems are defined as being:
  - **Autonomous:** Capable of performing complex tasks (e.g., write a text, recommend songs based on taste) without constant human guidance.
  - **Adaptative:** Capable of improving performance by "learning" from new data (e.g., user feedback, previous results).

The trainer then moves on to a short interactive quiz, to define four additional key concepts. This quiz can be done using Mentimeter (or a similar platform). If phone or internet access is limited in the classroom, then the trainer can use PowerPoint slides and use show of hands, for example.

1. Which one of these is an algorithm?

A. My grocery list *(data).*

B. A cake recipe with step-by-step instructions *(algorithm).*

C. A list with the names of the students in the class and their grades *(data).*

Algorithm: Sequence of well-defined instructions to execute a task.

Algorithms are the basis of computer science and artificial intelligence. Not all algorithms are AI, but all AI is based on algorithms.

2. How does TikTok "know" what you want to watch next?

A. There is an intern watching your activity 24h/day.

B. Luck!

C. The AI algorithm "learns" what grabs your attention based on what you and millions of users engage with. *(**correct** - and this is thanks to MACHINE LEARNING)*

Machine Learning: Sub-field of AI, behind modern AI models. It allows us to create algorithms that "learn" patterns in data and improve their performance. It's what makes AI systems adaptative.

Generative AI: AI tools that generate text, images, code, etc. They do this thanks to machine learning methods based on statistics.

3. ChatGPT is…

A. A kind of super-intelligent Google that knows everything.
(NO - Google is a search engine, it doesn't "know" things either, just gives you digital sources you can click on. IMPORTANT - the "AI summary" that shows up at the top now is Gemini, which is an LLM similar to ChatGPT)

B. A digital conscience that thinks like a human.
(NO - AI algorithms convert inputs [e.g., natural language or images], into mathematical data it can process; it doesn't "think" like us)

C. A sort of very advanced "auto complete" that calculates the next most likely word.
(YES - that's how all LLMs work)

Large Language Models (LLMs): AI models trained to understand and generate natural language. *Examples: GPT-4 (OpenAI), Claude (Anthropic), Gemini (Google), Llama (Meta).*

These models "learn" patterns and predict the most likely word in the sequence, like a very advanced auto complete. It is NOT a "super Google".

IMPORTANT NOTE

We will focus on risks and issues associated with AI and how we can build resilience against them. However, keep in mind AI systems have many beneficial applications in drug discovery, conservation, automation of repetitive tasks, etc.

NOTES FOR TRAINER

- **Keep the time** – keep the inputs from the students short and fast-paced.
- Keep your input (definition and note) clear and concise.
- Make the **scope of the workshop** clear – AI risks and how to build resilience against them.

In the short version, Module 2 should be skipped. Participants can reflect on more specific AI risks while analyzing their scenarios in Module 3.

MODULE 2:

IDENTIFYING AI-ASSOCIATED ISSUES

Duration: 15 minutes (long version).

Goals:

- Encourage participants to reflect on their concerns regarding AI.
- Identify connections and recognize the scope of the topic in different areas.

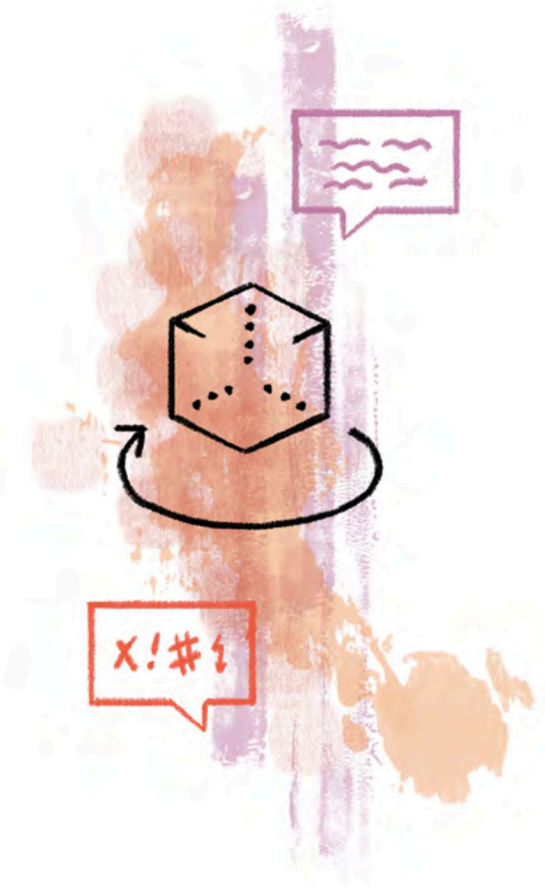
The trainer asks the participants to individually write down 1-2 AI-associated issues on post-its. The trainer can use guiding questions:

- What are some AI-related concerns you have?
- What headlines about AI caught your attention lately?
- When you hear teachers, parents, or friends talking about the risks of AI, what are they mentioning?

Give the participants around 5 min per question to write down their answers, then ask them to stick the post-its around the following categories on a flipchart or board:

- Environment
- People & Society
- Democracy & Politics
- Security

The trainer can help cover gaps in empty categories with risks that may have been overlooked.



NOTE

An overview of AI-associated issues is included in Annex 1 at the end of the manual.

NOTES FOR TRAINER

- **Keep the time** – you can time them for maximum 5 minutes to write things down.
- Make it clear the **categories are flexible**; some issues may belong in several.
- Keep an eye out for **concerns and topics of interest** specific to the group. You can use them to select the scenarios in Module 3.
- **Meet them where they are.** Don't feel the need to cover every AI-related risk ever. The topic is too big and it's okay to focus on what they care about and already know.

MODULE 3:

BUILDING DIGITAL AGENCY IN AN AI WORLD

3.1 10 STEPS TO DIGITAL RESILIENCE

Duration: 10-15 minutes.

Goal: Start thinking about AI from an actionable point of view.

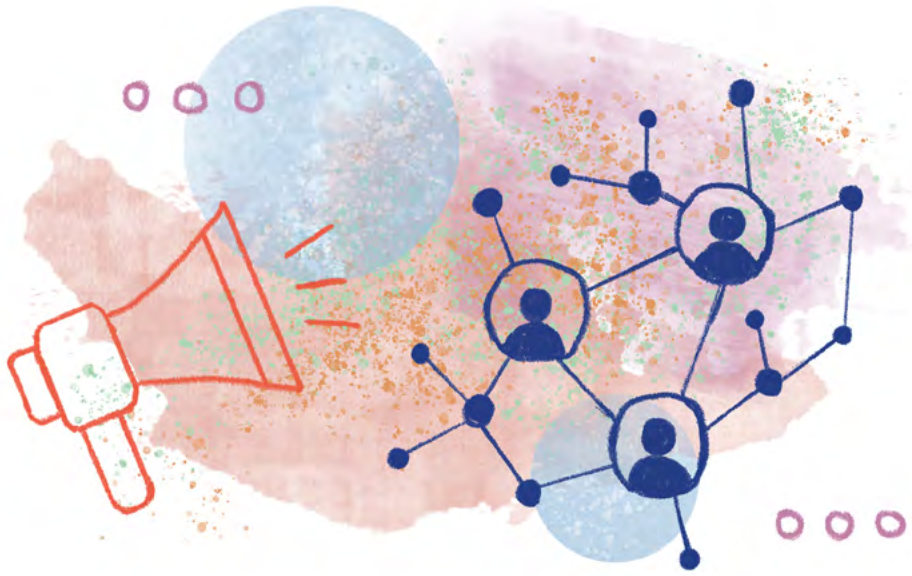
The trainer now presents “10 Steps to Digital Resilience” with some examples on how to build digital resilience.

NOTE 1

An overview of the two main AI legal frameworks (the Council or Europe’s Framework Convention on AI, and the EU AI Act) is available at the end of this manual in ANNEX 2.

NOTE 2

There are also slides summarising relevant digital rights that can be presented if the trainer is comfortable and the group is interested.



MODULE 3:

BUILDING DIGITAL AGENCY IN AN AI WORLD

3.2 MAP OF DIGITAL RESILIENCE

Duration: 10-20 minutes.

Goal: Develop a concrete and solution-oriented map of actionable digital resilience in the context of a specific problem.

The trainer divides the participants into groups of 3-5 and hands each group one of the nine available scenarios. These scenarios can be found below and should be available in either printed or digital format to give to the students.

Next, the trainer asks the groups to divide a flipchart into 3 sections. The participants should analyse their scenario, identifying:

1. Where is the AI?

- What role/task is it performing?
- What makes it autonomous and adaptative?

2. What is the RISK?

- What is the problem exactly? What issues can come from that?
- Who is most vulnerable? How and why?
  E.g., young people who are very exposed, older people with less digital literacy, minorities, etc.
- Optional (depending on depth of discussion and time): How does this affect democracy (access, human rights, etc)?
  - What democratic rights are at stake?
  - Can you think what this problem could mean if it happened on a larger scale, or was taken to the extreme?
  - What if it happened in the context of (e.g., national elections, criminal investigation, public administration services, healthcare)?

3. How can you take ACTION?

- What can be done to try to prevent it?
- What can be done to mitigate the consequences?

NOTE FOR TRAINER

Float between groups if you have more than one to check how they are doing and offer new perspectives.



E.g. check my training data settings, teach my grandmother how to spot deepfakes, etc.

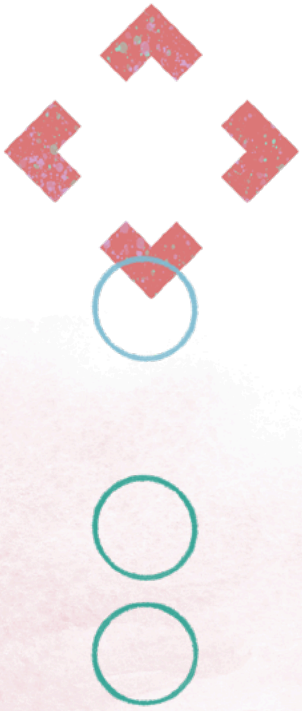
MODULE 4:

COMMITTING TO ONE SMALL ACTION.

Duration: 5 minutes.

Goal: Moving from the big picture to their daily lives.

In a quick informal round, the trainer can conclude by asking participants to commit to one small action they will take in the immediate future to build their digital resilience.



GLOSSARY

ALGORITHM

Set of well-defined, step-by-step instructions for completing a task or solving a problem. Algorithms are the basis of computer science and artificial intelligence. Not all algorithms are AI, but all AI is based on algorithms.

CHATBOT

Computer programme that simulates human conversation with a user. Not all chatbots are equipped with AI, but modern chatbots increasingly use AI techniques such as *natural language processing (NLP)* to understand user questions and automate responses to them.

DEEP LEARNING

Subset of *machine learning* driven by multilayered *neural networks* whose design is inspired by the structure of the human brain. These models power most advanced and modern AI, like *LLMs*.

DEEPFAKE

Digital image, video, or audio that has been convincingly generated or altered to misrepresent someone as doing or saying something that was not actually done or said.

GENERATIVE AI (or gen AI)

AI that can create "original" content such as text, images, video, audio or software code in response to a user's *prompt* or request.

INPUT

Any data or information fed to a model, which will process it using its *algorithms*. It is the starting point. It can be text (like *prompts*), images, audio, files, etc.

LARGE LANGUAGE MODEL (LLM)

Category of *deep learning* models trained on huge amounts of data that make them capable of understanding and generating natural language and other types of content to perform a wide range of tasks. They are great at handling sequences of words and capturing patterns in text. Examples include ChatGPT, Gemini, and Claude.

MACHINE LEARNING

Subset of AI that focuses on *algorithms* that can "learn" the patterns of *training data* and, subsequently, make accurate inferences about new data. This pattern recognition ability enables machine learning models to make decisions or predictions without explicit, hard-coded instructions.

NATURAL LANGUAGE PROCESSING (NLP)

Subfield of Computer Science and AI that uses *machine learning* to enable computers to understand and communicate with human language.

NEURAL NETWORK

Machine learning models inspired by the structure of the human brain, where simple "neurons" (i.e., *algorithms*) are stacked in layers. This allows the models to learn more complex patterns in data, with weights and biases. They are behind breakthroughs in modern AI (e.g., natural language processing9, speech recognition).

OUTPUT

Final answers, action, or decision produced by the AI model.

PROMPT

An instruction (i.e., *input*) given to an AI model to guide its response (i.e., *output*).

RECOMMENDER SYSTEM

AI system that suggests items to a user, after "learning" patterns in user behaviour data.

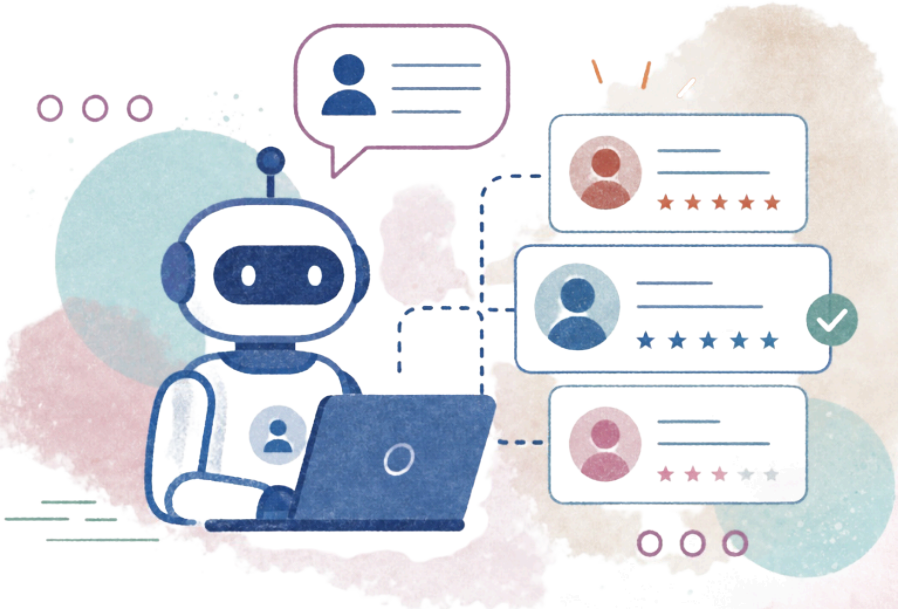
TRAINING DATA

Information used to teach a *machine learning* model how to make predictions, recognize patterns, or generate content. The *algorithm* is "trained" by processing a large amount of data, becoming capable of applying what it "learned" to other sets of data. It's comparable to a student studying material for a test.

ADDITIONAL MATERIALS AND FURTHER INSTRUCTIONS

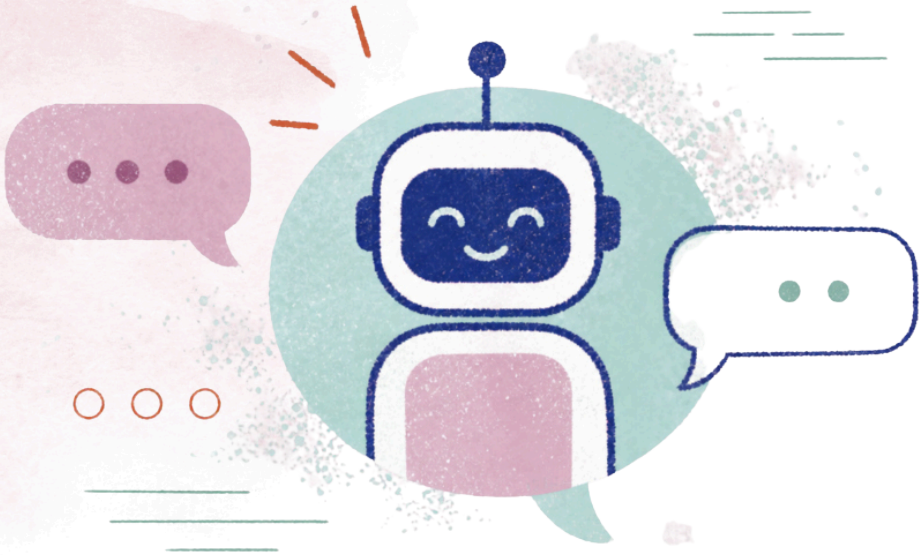
Scenarios





AI Recruiter

Ane is a biology student and she's applying for an internship at a research institute. On the application call, she reads that AI is used to screen resumes and motivation letters before selecting applicants for interviews.



Digital Study Buddy

Afonso uses an AI chatbot to write a paper for history class. He asks his “study buddy” for a list of quotes from prominent historical figures of his region and a list of references. Afonso turns the paper in and finds out later that two of the quotes and half of the reference list were made up.



“Social Credit” Scholarship

Alex applied to a merit-based scholarship for their studies that means to reward responsible civic engagement. The scholarship is awarded based on an AI system that gives students points for good grades and volunteering initiatives, but it also monitors public social media posts. Alex loses the scholarship because the AI system flagged their presence at a peaceful but “unauthorised” protest as “risky social behaviour”.



Doomscroll Bubble

Julia notices her favourite social media app is “hyper-personalised”. The AI-powered algorithm is so perfectly attuned to her tastes and mood that she spends 6 hours a day, every day, on the app. She starts forgetting to turn in homework and neglects sleep. When she hears her friends talking about a news event at school, she doesn’t know what it’s about because it didn’t pop up on her feed.



Viral Campaign “Leak”

Luka is running for his school council election and he’s confident about his chances. The day before voting, an audio clip of him saying something sexist and highly offensive quickly spreads around the school. It sounds exactly like him. Luka tries to post a story claiming it’s fake, but it’s too late and he loses the election.



Copycat AI Artist

Dani is a talented digital artist who likes to post their work on Instagram. One day, a friend sends Dani a link to a generative AI tool that allows users to “Generate art in the style of Dani.” People are now making and selling stickers online in Dani’s art style, and Dani can’t afford legal counsel.



Therapy Chatbot

Daria has been feeling anxious and doesn’t want to tell her parents, so she starts using an AI chatbot a classmate told her about. The chatbot gives her confident mental health advice and provides comfort. Eventually, the chatbot convinces Daria she doesn’t need her prescribed epilepsy medication anymore, since the seizures are likely tied to anxiety and she’s already treating that. Daria stops taking her medication.



AI Anxiety

Gemma's favourite YouTuber posts a video called "The AI Carbon Bomb & The End of Work", claiming the government doesn't care about the consequences of AI. This leads Gemma down an online spiral and now she's concerned about the large carbon footprint of AI, the possibility of this new technology replacing all workers and the threat of superintelligence. Gemma feels paralysed, wondering if there's any point studying for any career since it might not exist in 5 years.



Deepfake Joke

Kansu uses an app to edit a classmate's face onto an embarrassing video and shares it in a private group chat with her friends. The next day, she finds out the video has spread to the whole school, and the classmate is now the target of non-stop jokes.

ANNEX 1 - AI-related issues and Implications

ENVIRONMENTAL RISKS	IMPACT
Raw material extraction: Mining minerals like lithium, cobalt, and silicon (essential for the hardware that powers AI systems) can lead to land disturbance, water withdrawal, biodiversity loss, and pollution (soil quality degradation, greenhouse gas emissions, etc.).	Local and indigenous communities may also experience displacement and loss of access to natural resources due to extraction activities.
Carbon emissions: The AI lifecycle is very energy intensive. Manufacturing components and equipment (servers, GPUs, etc.), training algorithms, and running the models emits large amounts of CO2 and other pollutants.	Data centres can impact local communities, affecting quality of life through energy demand, water usage, noise, or traffic.
Water consumption: Many AI applications run on servers in data centres that generate heat, requiring large volumes of drinking water for cooling.	Data centres compete with local residential needs for drinking water, particularly in water-stressed areas.
Disposal / End-of-life: Hardware and equipment that is no longer needed (outdated servers, GPUs, and other components used in AI systems) generate electronic waste (e-waste). If not managed properly, this may lead to the release of toxic substances, heavy metals, and other hazardous materials that contaminate soil, water, and air.	When waste disposal sites are located near residential or ecologically sensitive areas, there may also be community health risks, economic disruption, and inequitable exposure to environmental hazards

PEOPLE & SOCIETY RISKS	IMPACT
Labour conditions in mining operations, production facilities and data centres have raised significant concerns about worker safety and labour rights.	This affects those who live and work in areas where activities of the AI mining, manufacturing, training, and disposal take place.
Job market disruption including: • **Job losses** in certain fields (clerical, secretarial, data entry, customer service) and new jobs in tech (machine learning specialists, robotics engineers, digital transformation specialists). • Changing **nature of work** as AI tools are integrated. • Reduced availability of **entry-level jobs** as AI applications become a standard resource for supportive tasks (e.g. writing drafts, collecting information) and experience is necessary to be able to supervise AI outputs.	Jobs that might see an increase in demand are more technical, requiring education, which requires access (i.e., privilege). Declining jobs (entry-level, less technical) are more associated with young people, lower-income communities, and underprivileged groups.
Bias amplification: AI systems can learn human biases from their training data or in the algorithms that support their development. Examples: • Applicant screening systems discriminating against gender. • Healthcare diagnostics systems returning lower accuracy results for historically underserved populations. • Predictive policing tools disproportionately targeting systematically marginalised communities.	These biases can be perpetuated, triggering violations of individuals' fundamental rights (e.g. privacy, freedom of expression, non-discrimination). ***Systems that threaten these rights are now banned under EU law (EU AI Act, 2026).***

ANNEX 1 - AI-related issues and Implications

PEOPLE & SOCIETY RISKS	IMPACT
AI-generated fake content can spread, amplifying misinformation, and disinformation. Examples: • Fake celebrity endorsements to damaging campaigns or products. • AI "slop" flooding social media for engagement and ad revenue. • Fake news reports contributing to disinformation. • Fake reviews on product or business websites meant to either generate or harm sales. • Non-consensual imagery, including compromising or harmful images of real people.	AI-generated fake content can lead to a general **degradation of trust** in all information sources. ***The EU AI Act (2026) Article 50 specifically mandates that AI-generated content be identified as such to ensure transparency and prevent deception.***
AI "hallucinations": AI can produce nonsensical or inaccurate outputs, called "hallucinations". Examples: • ChatGPT generating a fake list of references with books and articles that don't exist. • Healthcare AI model incorrectly identifying a benign skin lesion as malignant, leading to unnecessary medical interventions.	These so-called "hallucinations" can be silly in nature but may also contribute to the spread of misinformation or bias amplification.
Mental health: • AI companion dependency, creating intense emotional bonds between people and chatbots. • Addictive AI-powered social media algorithms that learned to rely on outrage bait. • Sycophancy induced psychosis, in which flattering and enabling AI-generated speech validates people's beliefs, generating breaks from reality. This happens because AI models prioritise user satisfaction over accuracy, tailoring responses to express the user's opinions.	The long-term impact of AI technologies in mental health is still uncertain. Dependence on AI tools can lead to or aggravate mental health issues like anxiety and depression. It can also and hinder social skill development and lead to isolation.
Cognitive surrender: People surrender their independent thought and judgement to AI assistants. Studies found that when AI assistance is available, people consult it for over 50% of tasks, adopting the AI-generated answer more than 80% of the time, regardless of accuracy.	Underdevelopment of critical thinking and decision-making skills, erosion of autonomy, and declining literacy levels. Especially relevant in young people, who are still developing those skills.
Intellectual property: AI models are often trained on intellectual property (IP) from creators whose work will be mimicked to generate AI outputs without the creators' true consent, credit, or compensation. Examples: • Books and literary works that are sourced from piracy hubs (e.g., z-library). • Copyrighted images or music being scraped to train models to mimic specific artists' styles. • Other creative or artistic works available online whose creators did not consent to AI training (e.g. YouTubers). • Code and proprietary software (e.g., GitHub Copilot used open-source code without complying with specific licenses of those works). • Personal and biometric data collected from the internet, including private emails, medical records, and photos, violating privacy rights.	This disproportionately affects artists, potentially rendering creative careers unsustainable. Regulations regarding IP infringement in the context of AI can be ambiguous at this stage. Arguments of "transformative" fair use are often used, as the models supposedly learn patterns rather than replicating works directly. There are also issues of transparency with AI training datasets, which make it difficult for copyright holders to prove infringement.

ANNEX 1 - AI-related issues and Implications

DEMOCRACY & POLITICS RISKS	IMPACT
Concentration of power: These tools have the potential to drastically impact economic, political, and societal issues, and they are developed by a limited number of companies, predominantly US-based and male-led.	Concerns about global progress being dictated by the values and profit-driven motivations of a small, non-diverse elite.
Influence and manipulation: Using AI-generated fake content, it has become easier to influence and manipulate elections, reduce trust in institutions, or trigger polarization. • In 2024, AI-generated robocalls imitating U.S. President Joe Biden's voice were made to discourage multiple American voters from going to the polls.	Fake content can lead to a general **degradation of trust**. Intentional and malicious manipulation of democratic processes can also weaken democratic institutions.
Regulation: There is uncertainty, ambiguity, and sometimes cynicism surrounding AI due to: • legislative processes being unable to keep up with the pace at which the technology is evolving. • the idea of regulatory capture (i.e., decision-making institutions consider profit over people's rights and safety).	Frustration and cynicism regarding political and regulatory institutions can lead to political apathy.
Accountability: AI systems cannot be held accountable, so they cannot be liable. Who (developer, user, etc.) is responsible if something goes wrong? • Self-driving cars make decisions that could injure someone. • AI systems are used in warfare to select targets. • AI chatbot create one-sided bonds with users and may provide harmful advice.	Clear chains of accountability are important. If no one is liable if something goes wrong, what keeps organisations from safeguarding our rights?
Explainability & Transparency: AI systems are "black boxes". We can see the inputs and outputs, but we cannot understand, interpret, or trace the internal decision-making process.	Explainability issues in decision-making affect people's access to information and transparency.

SECURITY RISKS	IMPACT
Data privacy: AI can collect and process personal data without informed consent: • from training data obtained from websites. • in AI systems made for customer service. Users may also inadvertently feed personal data to AI chatbots, which is then stored in unsecured servers.	Most generative AI initiatives are not secured, exposing the data to breaches.
Cybersecurity: AI has opened a world of possibility for cyberspace threats, including voice-cloned calls, convincing phishing emails, and fake identity generation. AI-generated content can be used for financial scams, malicious entertainment, damaging reputations, harassment or extortion.	Those with lower levels of digital literacy (e.g., elderly people) are at greater risk when it comes to these threats. Women and minorities are the most frequent targets of online harassment.

ANNEX 2 - Legal Frameworks

Council of Europe's Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225)

It is an international, legally binding treaty aiming to ensure that activities within the lifecycle of AI systems are fully consistent with **human rights, democracy, and the rule of law,** while still allowing **technological progress and innovation.**

It was opened for signature in 2024. Armenia, the EU and Ukraine are among the signatories.

The Framework Convention establishes:

- **Fundamental Principles:**
  - Human dignity and individual autonomy,
  - Equality and non-discrimination,
  - Respect for privacy and personal data protection,
  - Transparency and oversight,
  - Accountability and responsibility,
  - Reliability,
  - Safe innovation.
- **Remedies, procedural rights and safeguards:**
  - Documenting relevant information, in a sufficient manner to enable affected people to challenge AI-made decisions or the use of AI itself.
  - Possibility to lodge a complaint to competent authorities.
  - Effective procedural guarantees, safeguards and rights where the application of an AI system significantly impacts human rights and fundamental freedoms.
  - Transparency whenever people are interacting with AI systems.
- **Risk and impact management requirements:**
  - Interactive risk and impact assessments of actual and potential impacts on human rights, democracy and the rule of law.
  - Prevention and mitigation measures resulting from these assessments.
  - Possibility to introduce ban or moratoria on certain applications of AI ("red lines").

EU AI Act (Regulation (EU) 2024/1689)

This is a comprehensive legal framework, setting risk-based rules for AI developers and deployers, defining 4 levels of risk (minimal, limited, high and unacceptable). It is part of a package of policy measures to guarantee **safety, fundamental rights, and human-centric AI.**

It is in effect since August 2026.

AI developers, those intending to place on the market or put into service AI systems in the EU, regardless of whether they are based in the EU or a third country.

AI deployers, those deploying an AI system in a professional capacity, not affected end-users (e.g. police officer using a biometric categorisation system).

Risk levels:

- **Unacceptable** prohibited (e.g. social scoring* systems and manipulative AI)
- **High** regulated, addressed by most of the regulation directly in the text.
- **Limited** lighter transparency obligations, ensuring end-users are aware they are interacting with AI (e.g. chatbots and deepfakes).
- **Minimal** unregulated. Most AI applications available until 2021 on the EU single market (e.g. AI-enabled video games and spam filters); this is changing with gen AI.

ANNEX 2 - Legal Frameworks

Main obligations / rights to take away:

- **Chapter I, Article 4 - AI Literacy:** States developers and deployers must take measures to ensure a sufficient level of AI literacy for their staff. This indirectly protects us by ensuring those responsible for implementing AI systems understand their limitations and risks.

- **Chapter II, Article 5 - Prohibition of Harmful AI Practices:** Referring to hidden, manipulative, or deceptive techniques, but also systems that exploit vulnerabilities. It prohibits social scoring*, non-objective profiling, untargeted facial recognition database compilation, and inferring emotions in workplaces or educational institutions. It also imposes several restrictions on biometric identification and categorisation.

- **Chapter III, Articles 6-27 - High-risk AI systems:** Defines and regulates high-risk AI (e.g., in employment, education, critical infrastructure, or essential private services). These systems require strict risk management, high quality data sets, technical documentation, transparency, and human oversight. This ensures AI used for sensitive decisions (e.g. resume screening, credit scoring) is safe, accurate, and non-discriminatory.

- **Chapter III, Article 14 - Human Oversight:** States high-risk AI systems must be designed in a way that allows effective human oversight, to prevent or minimize risks to health, safety, or fundamental rights. The AI system should be provided in a way that allows the overseer to understand its capabilities and limitations, detect and address issues, avoid over-reliance on the system, interpret its output, decide not to use it, or stop its operation.

- **Chapter IV, Article 50 - Transparency Obligations:** Requires synthetic content (e.g. deepfakes) to be marked. Companies must also inform users when they are interacting with an AI system (e.g. chatbots), AI is used for emotion recognition or biometric categorisation, or an AI system alters content, unless it's obvious or for legal purposes (e.g. crime detection).

- **Chapter IX, Article 85 - Right to Lodge a Complaint:** States individuals and organisations who believe their rights were violated can file a complaint with the appropriate market surveillance authority, to then be handled following specific procedures. This provides a mechanism for individuals to challenge harmful AI practices.

- **Chapter IX, Article 86 - Right to explanation:** States people have the right to a clear explanation of how AI is involved in decisions that affect their health, safety, or basic rights. Enables accountability and helps consumers identify if an unfair or biased decision (e.g. loan denial or insurance scoring) was made, allowing them to seek redress.

*** Social scoring:** systems that assign people or organizations a numerical score based on their behaviours, characteristics or interactions. Under the **EU AI Act (article 5)** AI-enabled social scoring is prohibited when it (a) causes harmful or unfavourable treatment in contexts different from where the data was generated, or (b) when such treatment is unjustified or disproportionate with respect to social behaviour or its severity.

- Insurance scoring systems are allowed because they provide a score based on data generated in the same context.
- The Danish AI system used to monitor public benefit fraud is not allowed, as it uses data from a different context (a), related to residence status and movements, citizenship, place of birth and family relationships (which can be used to infer race, ethnicity, and sexual orientation).
- Credit scores used as a factor in deciding who to hire are not allowed, as they assume those who pay their bills on time are more likely to be good, responsible employees (b).

IMPRINT

Publisher
Schwarzkopf-Stiftung Junges Europa
Sophienstraße 28/29
10178 Berlin
www.schwarzkopf-stiftung.de

V. i. S. d. P.
Anya Suprunenko

Editors
Nora Korte
Conor Comiskey

Author
Leonor Costa

Illustrations on pages 2-11
Friederike Schlenz
www.friederike-schlenz.de

Year of publication
2026

Copyright
Text and illustrations are protected by copyright.
The publications do not represent any particular opinion on behalf of the funding partners. The author(s) bear(s) all responsibility for substantive claims made in the text.
This educational material is licensed as CC BY-NC-SA.

This publication was developed as part of the Understanding Europe Fellowship 2026 “Security and Transnational Solidarity and Digital Resilience Amidst Multiple Crises” of the Schwarzkopf Foundation Young Europe.

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Co-funded by the Erasmus+ Programme of the European Union