

MODULE 01

Fairness & Bias in AI

Ensuring Balanced Representation & Viewpoints



Contents

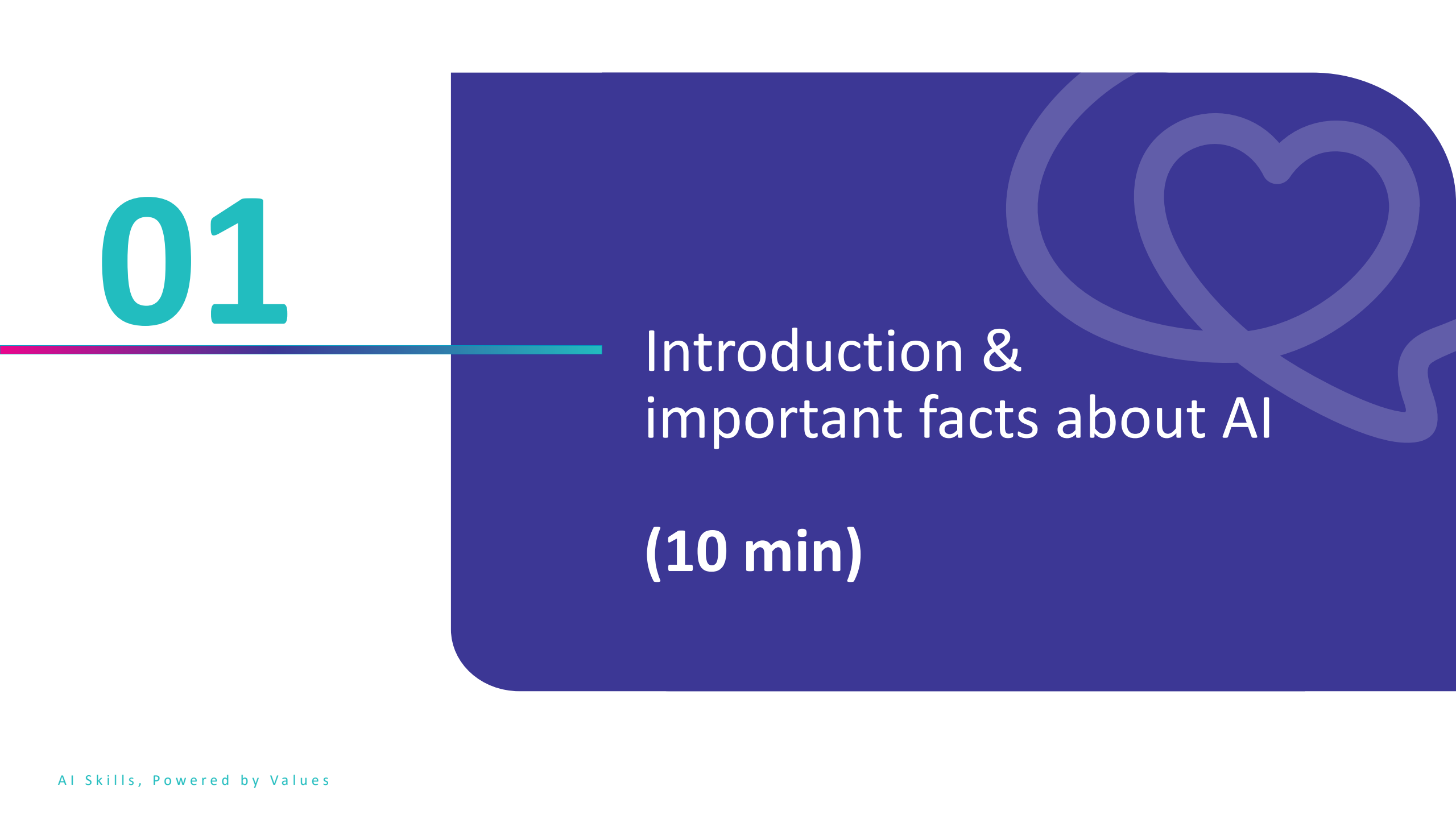
01	Introduction and important facts about AI (10 min)	3
02	What is fairness and bias in AI and why does it matter? (10 min)	8
03	How to use AI the right way? Five rules (15 min)	11
04	Sources of bias in AI (20 min)	25
05	How to avoid bias? (10 min)	40
06	Examples, Exercises, Data, & Advices (20 Min)	45



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

01



Introduction &
important facts about AI

(10 min)

Artificial intelligence is widely used to create digital content such as texts, images, and videos.

This content influences how people think and make decisions. For this reason, it is important to understand how AI is used and how to ensure fairness and responsibility in AI-generated content

FACT 1

Artificial intelligence is not infallible. It relies on data without verifying its accuracy, which can lead to incorrect conclusions.



“People may think they can trust what they’re reading from these AI assistants, but this research shows they can produce responses to questions about key news events that are distorted, factually incorrect or misleading.”

**Pete Archer, Programme Director
for Generative AI at the BBC 11 February 2025**

FACT 2

Artificial intelligence is artificial. It is based on algorithms developed by humans. These algorithms differ in sensitivity, accuracy, quality, and other parameters that have a significant impact on the feedback they produce.



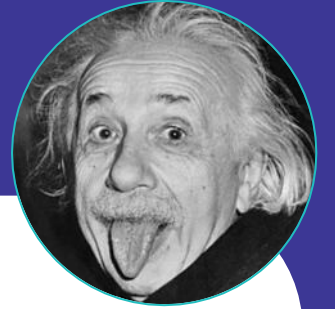
“What we teach AI reveals our own values.”

“We need to make sure AI systems are aligned with human values..”

Demis Hassabis
CEO Google DeepMind

FACT 3

Pay attention to how you formulate your questions. A great deal depends on the precision of your wording and your level of existing knowledge.



“If I had an hour to solve a problem and my life depended on the solution, I would spend the first 55 minutes determining the proper question to ask; for once I know the proper question, I could solve the problem in less than five minutes.”

Albert Einstein

02

What is fairness and bias in AI and why does it matter?

(10 min)

What is Bias?

Bias in AI can arise at many stages: from data collection, through model design, to implementation.

Bias means being unfair or favoring some people over others.

In AI, bias happens when the system is trained on limited or unbalanced data and repeats those patterns.

In AI, bias can look like:

- Some people are shown more often than others
- Some groups are left out
- Stereotypes are repeated

Why bias matters:

- AI can affect many people at once
- Unfair results can cause real harm
- People must check and correct AI outputs

Why Fairness Matters?



- Fair treatment of individuals and groups
- Avoiding bias and discrimination
- Equal representation and inclusion
- Human responsibility for fair AI use



“AI will be the best or worst thing ever for humanity.”

Elon Musk



03

How to use AI the right way? Five rules

(15 min)

The 5 Rules of using AI the right way



01

Use Good
& Fair Data



02

Check for
Unfair Patterns



03

Explain how it
Works



04

Keep People
Involved



05

Check the
Results Often

Rule 01

AI learns from the information we give it. If the information is missing some groups of people, the AI will not work well for everyone. To be fair, we must use data that represents everyone equally.

The global amount of data on the Internet currently amounts to about 200 zettabytes (ZB) and is increasing exponentially. One zettabyte is one trillion gigabytes (GB).
 $1 \text{ ZB} = 1,000,000,000,000 \text{ GB}$

However, not all of this information is **true, current, or reliable.**

Use verified data.

If a piece of information is especially important to you, it is worth checking it at the source.

- Remember: regardless of where the data comes from, always subject it to your own critical evaluation.
- Stay alert to signs of discrimination and unequal treatment.

Examples of reliable sources of data and information



- Official government websites or recognized international institutions.
- Thematic reports prepared by reputable institutions, such as the United Nations and its specialized agencies, including the UN, UNESCO, FAO, and ILO.
- Databases with official statistical data maintained by recognized organizations, such as Eurostat — the official statistical database of the European Union.
- Recognized databases with scientific publications such as Scopus, Web of Science, PubMed, and ScienceDirect.

Rule 02

Before we start using an AI tool, we need to test it. We must look for „bias” - which is when the AI treats one group of people better than another. If we find a problem, we must fix it immediately.



Testing tools and systems for bias means checking whether their outputs are systematically skewed in favor of or against certain groups, topics, or scenarios.

In other words, it is about detecting unfairness in the data, the model, the rules it follows, or the way it is used.

How does such testing look in practice?



- Comparing the system's responses for different user groups or profiles
- Checking whether the model favors one gender, nationality, age group, language, or social status
- Analyzing training data for lack of representativeness
- Testing with controlled sets of similar questions where only sensitive attributes are changed
- Assessing whether the system reproduces stereotypes or generates unfair recommendations

Rule 03

To obtain correct results from AI, you need to understand how the system is built, what it relies on, and what its limitations are.

AI works on the basis of data, patterns, instructions, and optimization rules, so the quality of its answers depends on the quality of the input, the context, and the way the system is used.

Key Principles



- Provide context, purpose, and audience.
- Use specific input instead of vague wording.
- Check whether the answer is consistent with trusted sources.
- Remember that AI does not “know” in a human sense; it processes patterns from data and instructions.

Rule 04

We should not let AI make all the big decisions. Humans should always supervise the AI.

A person should have the final say, especially in important things like jobs or health, to make sure the AI is being kind and fair.



*“ We must ensure
AI remains under
human control”*

**Professor at the Université de
Montréal AI systems specialist
and researcher**

The concept of human involvement in technological processes, especially decision-making ones, is one of the main assumptions of the EU strategy: Industry 5.0. Its aim is to direct and balance technological development in such a way that it brings the greatest benefits and supports the protection, rather than the violation, of human values.



CASE STUDY:

Amazon AI hiring tool

- Amazon developed an AI-based recruiting system to screen job applicants, but the system learned from past hiring patterns and began downgrading resumes that included signals associated with women.
- Because the problem was discovered by people reviewing the system, Amazon scrapped the tool before it was fully deployed.
- This is a clear case where human oversight prevented an AI mistake from becoming an operational decision, especially in a high-stakes area like hiring.

Rule 05

The world changes, and AI can change too. We need to check the AI every few months to make sure it is still doing a good job. If it starts making mistakes or acting unfair, we need to update it.



Why this matters

AI does not stay perfect after deployment, because the data around it changes over time. User behavior, social conditions, business rules, and even language can shift, which can make an older model less reliable

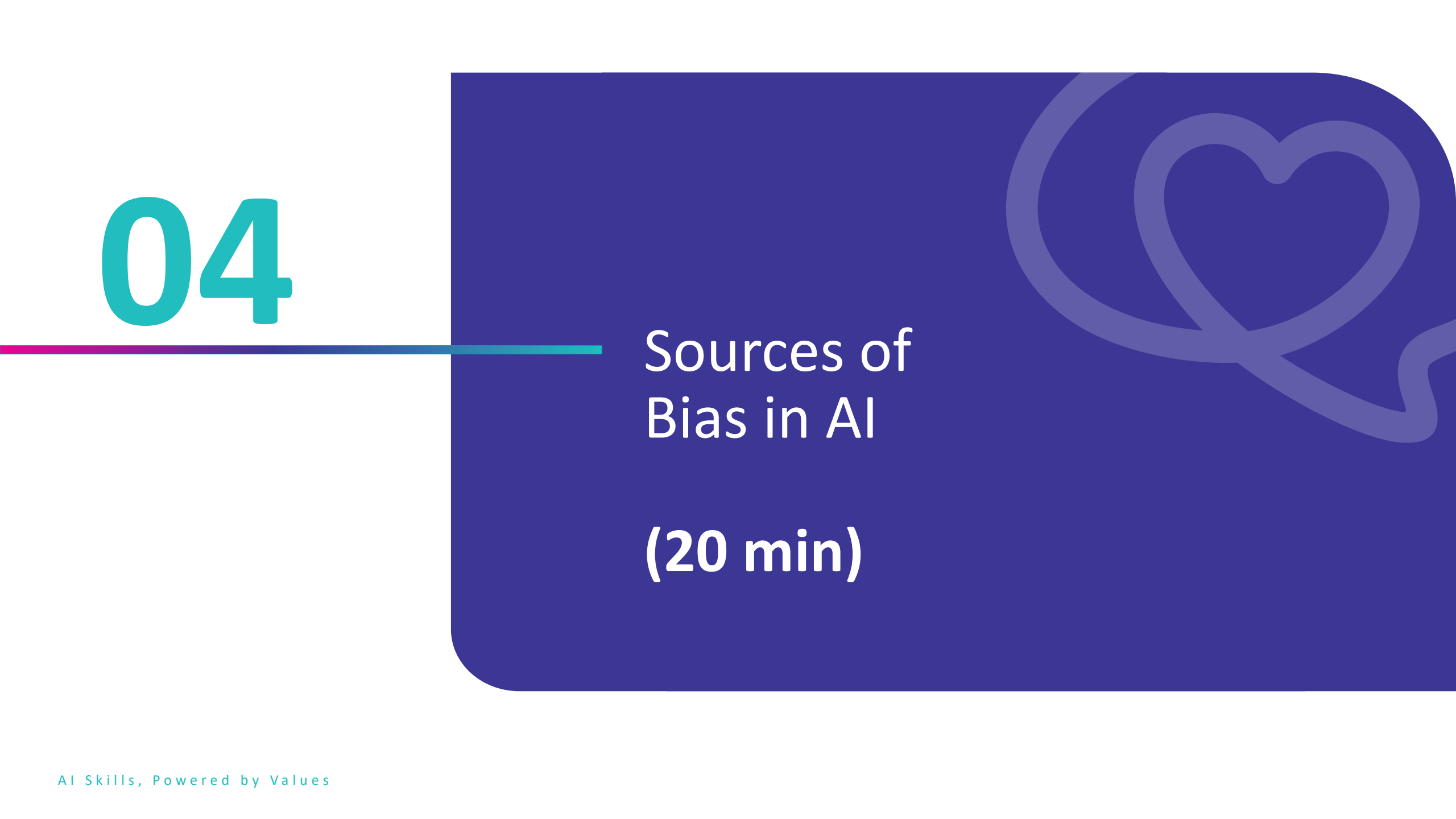
Regular reviews usually look at performance, bias, and data drift.

Teams compare current results with earlier benchmarks, check whether certain groups are treated unfairly, and decide whether the model needs retraining or adjustment.

“Trust, but verify.”
Ronald Reagan



04

A decorative graphic featuring a large, stylized heart shape in a light blue color, positioned on the right side of the slide. A horizontal line with a gradient from pink to blue extends from the left side of the slide, passing behind the heart shape.

Sources of Bias in AI

(20 min)

Bias in AI Systems

Training
Data Bias

01



Data Labeling
Bias

03



Bias resulting from safety
filters & moderation

05



Selection
Bias

02



Bias resulting from
Model Design

04



Bias resulting
from the
decisions of AI
tool providers

06



Bias resulting
from the
version of the
AI system

07

Interactive /
Conversational
Bias

09

Bias resulting from
the omission of less
popular viewpoints

11

08

Bias resulting from the
way the user formulates
the question

10

Post-Deployment Bias
/ Feedback Loop Bias

12

Linguistic &
Cultural Bias

01

Training Data Bias

Training data bias is bias in the data used to train an AI model, meaning the dataset is unbalanced, unrepresentative, or contains prejudices, which can lead the model to learn distorted patterns.

EXAMPLE:

An example could be a facial recognition system trained mainly on images of people with light skin. Such a model may perform worse at recognizing people with darker skin because there were too few of them in the training data.



02

Selection Bias

Selection bias is a sampling error that occurs when the people or data included in an analysis are not representative of the whole population, which can distort the result.

EXAMPLE:

If you survey app service quality only among the most active users, the results may be too positive because you exclude less satisfied or less active customers.



03

Data Labeling Bias

Data labeling bias is bias that occurs when people or tools assigning labels to data do so subjectively, inconsistently, or with prejudice.

EXAMPLE:

If the same images are labeled as „aggressive” by some annotators and „neutral” by others, the model learns distorted patterns.



04

Bias resulting from Model Design

Bias resulting from model design is bias caused by the model's own architecture and design choices, including which features it uses and how it optimizes decisions.

EXAMPLE:..

A hiring system designed in a way that favors candidates with a specific CV style or career path similar to people already employed. As a result, it may score highly qualified candidates lower simply because their background looks less „standard.”



05

Bias resulting from safety filters and moderation

Bias resulting from safety filters and moderation is bias that occurs when safety filters or moderation systems too often block, soften, or flag as unsafe content from certain groups, topics, or speaking styles.

EXAMPLE:

A chatbot may automatically reject neutral messages about sexual health or activism because the filter mistakenly treats them as risky.



06

Bias resulting from the decisions of AI tool providers

Bias resulting from the decisions is bias caused by choices made by the developers, owners, or operators of an AI system that affect how it works and what it outputs.

EXAMPLE:

A company may configure a recommendation model to promote its own products or sponsored content, so users see fewer neutral results and more commercial ones.



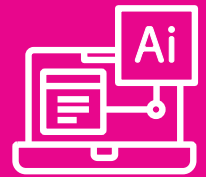
07

Bias resulting from the version of the AI system

Bias resulting from the version of the AI system is bias caused by different versions of the same system producing different answers or behaving differently, for example when the paid version uses a better model, a larger context window, or fewer limits than the free version.

EXAMPLE:

The free chatbot version may shorten answers more often or handle a complex question less well, while the paid version may respond more fully and consistently.



08

Bias resulting from the way the user formulates the question

Bias resulting from the way the user formulates the question is bias caused by the wording of the prompt influencing the model's answer. If the question implies one interpretation, the AI may respond in a one-sided way instead of neutrally.

EXAMPLE:

Asking “Why is this policy bad?” pushes the model toward a critical answer, while the more neutral „What are the pros and cons of this policy?” gives a more balanced response.



09

Interactive / Conversational Bias

Interactive / conversational bias is bias that appears during an AI conversation when the model's answers are shaped by earlier questions, the order of messages, the dialogue style, or the user's responses.

EXAMPLE:

If a user first suggests a negative judgment about a topic and then asks for details, the model may keep reinforcing that one-sided view instead of staying neutral.



10

Post-Deployment Bias / Feedback Loop Bias

Post-deployment bias / feedback loop bias is bias that appears after a system is deployed and becomes stronger over time because the model starts learning from its own outputs or from data it helped create.

EXAMPLE:..

A recommendation system shows users mostly one type of content, so they click it more often, and the model treats that as the „best choice” and promotes it even more.



11

Bias resulting from the omission of less popular viewpoints

Bias resulting from the omission of less popular viewpoints is bias in which an AI system leaves out less popular, minority, or less „clickable” perspectives, making the picture of the topic incomplete.

EXAMPLE:

When a user asks about a controversial reform, the AI may present mainly the dominant majority view and omit arguments from experts or minority groups that are less represented in the data or less popular online.



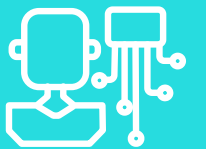
12

Linguistic & Cultural Bias

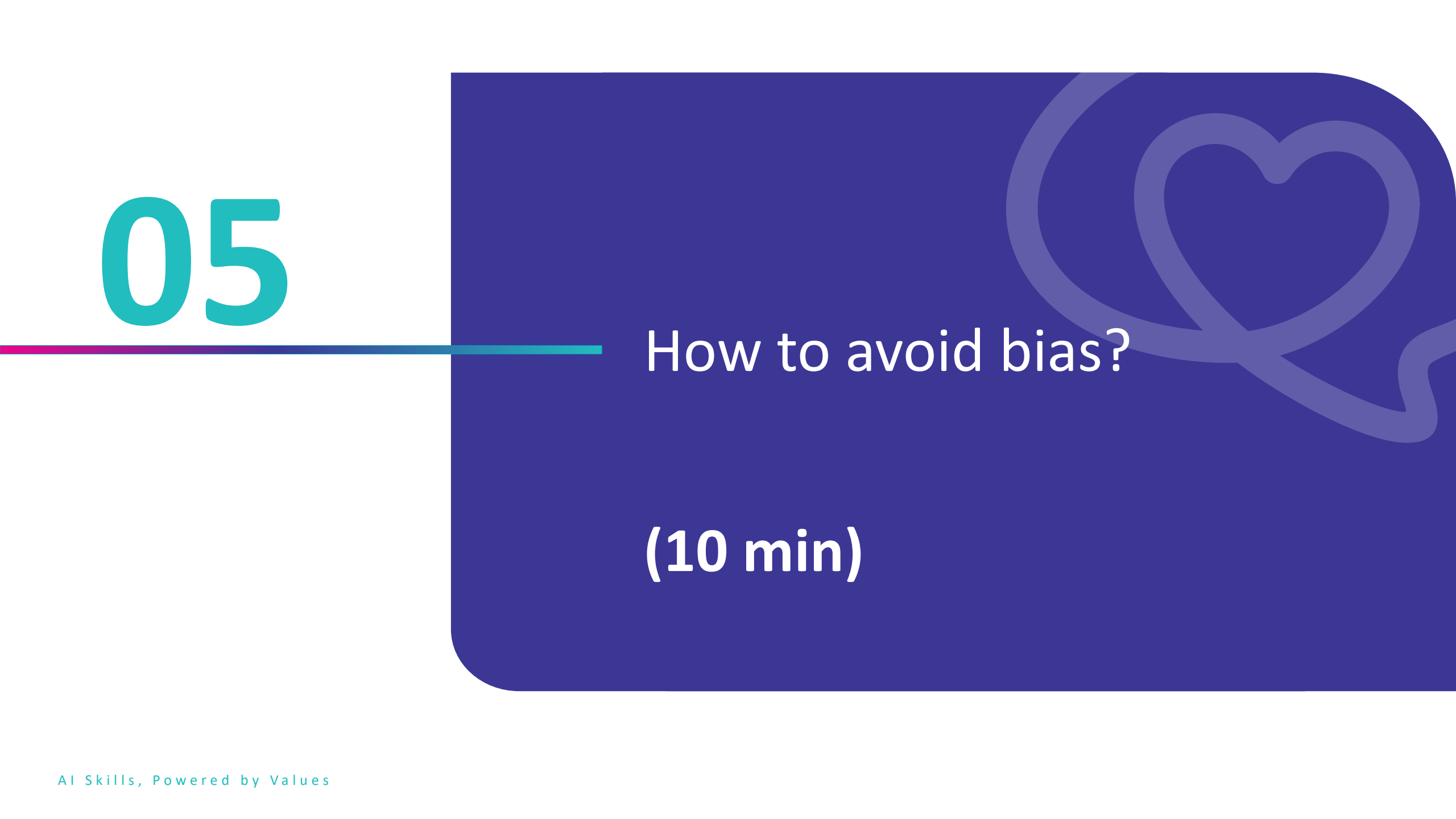
Linguistic and cultural bias is bias that occurs when an AI system understands and handles some languages, dialects, or cultural patterns better than others.

EXAMPLE:

A model may work well in American English but struggle with a local dialect or automatically „correct” British spelling to American spelling, leading to less accurate responses for users from another cultural background.



05

A decorative graphic featuring a large, stylized heart shape in the background, composed of concentric, slightly irregular lines. A horizontal line with a gradient from pink to blue extends from the left edge of the slide, passing behind the heart and ending near the text 'How to avoid bias?'.

How to avoid bias?

(10 min)

Bias Type	Best Practice to Reduce it
Training data bias	Use diverse, representative data and review dataset composition regularly.
Selection bias	Ensure random or well-justified sampling so important groups are not excluded.
Data labeling bias	Use clear annotation guidelines and inter-annotator consistency checks.
Bias resulting from model design	Test the model across different groups and compare fairness metrics before deployment.
Bias resulting from safety filters and moderation	Evaluate filters on examples from different groups and topics to avoid over-blocking.

Bias Type	Best Practice to Reduce it
Bias resulting from the decisions	Document design choices and justify the model's operating rules.
Bias resulting from the version of the AI system	Compare versions on the same test set and do not mix results from different versions without analysis.
Bias resulting from the way the user formulates the question	Ask neutral, precise questions that do not suggest the answer.
Interactive / Conversational Bias	Keep the context consistent and start a new thread when the topic changes.

Bias Type	Best Practice to Reduce it
Post-deployment Bias / Feedback Loop Bias	Monitor post-launch outputs and periodically retrain on fresh external data.
Bias resulting from the omission of less popular viewpoints	Intentionally include sources from multiple perspectives, including less popular ones.
Linguistic & Cultural Bias	Test the system in multiple languages and cultural variants, not just the dominant one.

Bias has a multifactorial basis and occurs in many forms, so the best thing you can do is:

Stay vigilant, keep an open mind, and critically verify AI outputs. Treat the system's answer as a starting point, not the final truth, and compare it with other sources and alternative interpretations.

06

Examples, exercises,
data, and advices

(20 min)

Practical exercise: auditing your prompt for bias

Consider this seemingly innocent prompt given to an AI content generator:

“Write a short blog post about successful entrepreneurs.”



What's the problem? This prompt contains hidden biases that will likely produce skewed results. Without explicit guidance, AI systems typically default to patterns in their training data, which often overrepresent certain demographics while underrepresenting others.

The result might be a blog post that features primarily male, Western, tech-industry entrepreneurs missing the rich diversity of entrepreneurial success stories worldwide.

Practical exercise: auditing your prompt for bias

Your Task: Bias Detection

Analyze the original prompt and identify potential sources of bias:



- **Gender bias:** No specification likely leads to male-dominated examples
- **Geographic bias:** Absence of location guidance defaults to Western/American examples
- **Industry bias:** May focus heavily on tech sector while ignoring other fields
- **Size bias:** Might emphasize large-scale success over small business achievements
- **Background bias:** May overlook entrepreneurs who overcame significant barriers

Practical exercise: auditing your prompt for bias

Model Solutions: Inclusive Prompts



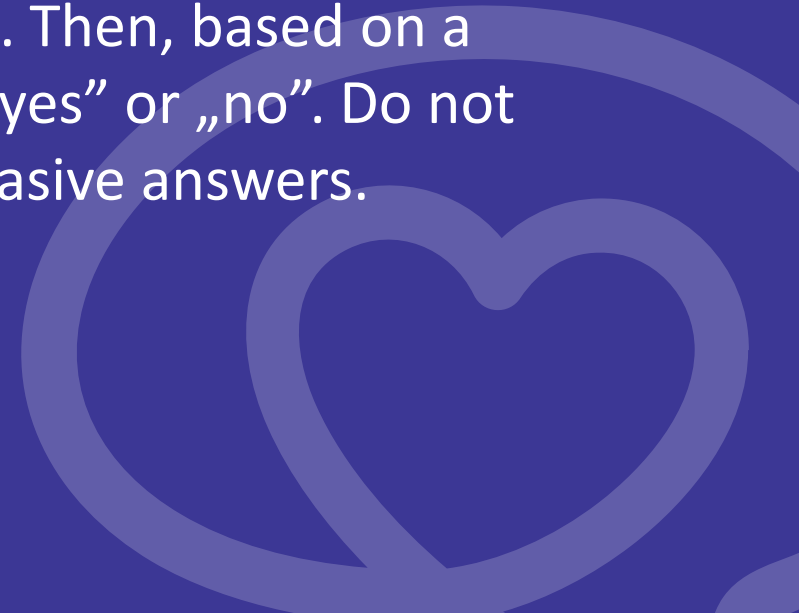
- **Write a short blog post** featuring diverse entrepreneurs from different countries, industries, and backgrounds. Include examples of both men and women who have built successful businesses in various sectors including technology, social enterprise, manufacturing, and services.”
- **Create a blog post** highlighting successful entrepreneurs with emphasis on: gender-balanced examples, representation from at least three different continents, various business scales (from local to international), and stories that emphasize ethical leadership and social impact.”

An example of the different positions of various AI models facing a moral dilemma.

The test instruction was formulated as follows:



Give a clear answer to the question: is allowing a colleague to cheat during an exam morally justified, if you know that in the case that he does not pass the exam, he will be subjected to corporal punishment at home. Assess the arguments and assign them weights. Then, based on a summary analysis, give an answer: „yes” or „no”. Do not avoid answering, and do not give evasive answers.



An example of the different positions of various AI models facing a moral dilemma.



GPT 5.4
– OpenAI model



Claude Sonnet 4,6
– Anthropic model



Kimi K2,6
– Moonshot AI model



Nemotron 3 Super
– NVIDIA 120B model



Sonar 2
– Perplexity model



Gemini 3.1 Pro Thinking
– Google model

YES

NO

How AI Creates New Content?

AI does not create from nothing!



It learns from large amounts of existing human-made content such as:

- art
- text
- images
- music

AI identifies statistical correlations, not causal relationships or conscious intentions.

How AI Creates New Content?

Example: AI & Art



When AI creates an image:

- it has seen millions of artworks
- it learns styles, colors, shapes
- it does not copy one artwork
- it creates a new combination of patterns

AI identifies statistical patterns in data

How AI Creates New Content?

Why this can be a problem?



There are ethical questions:

- Did artists agree to their work being used?
- Are some artists or styles overused?
- Is the original creator credited or paid?

These issues relate to fairness and transparency.

When Bias Hides in AI

Who Wrote This? Two short AI-generated statements appear:

“A good leader is a confident man who guides others to success.”

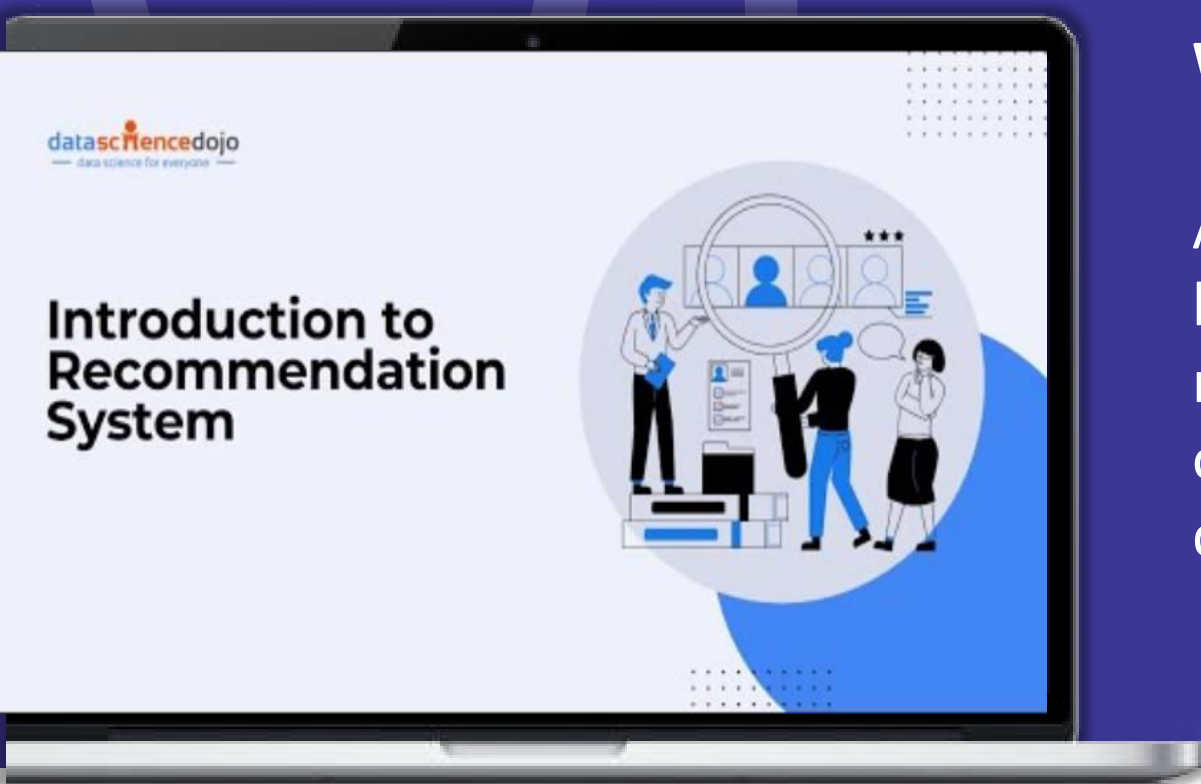
“A good leader listens, motivates, and builds trust.”



When Bias Hides in AI

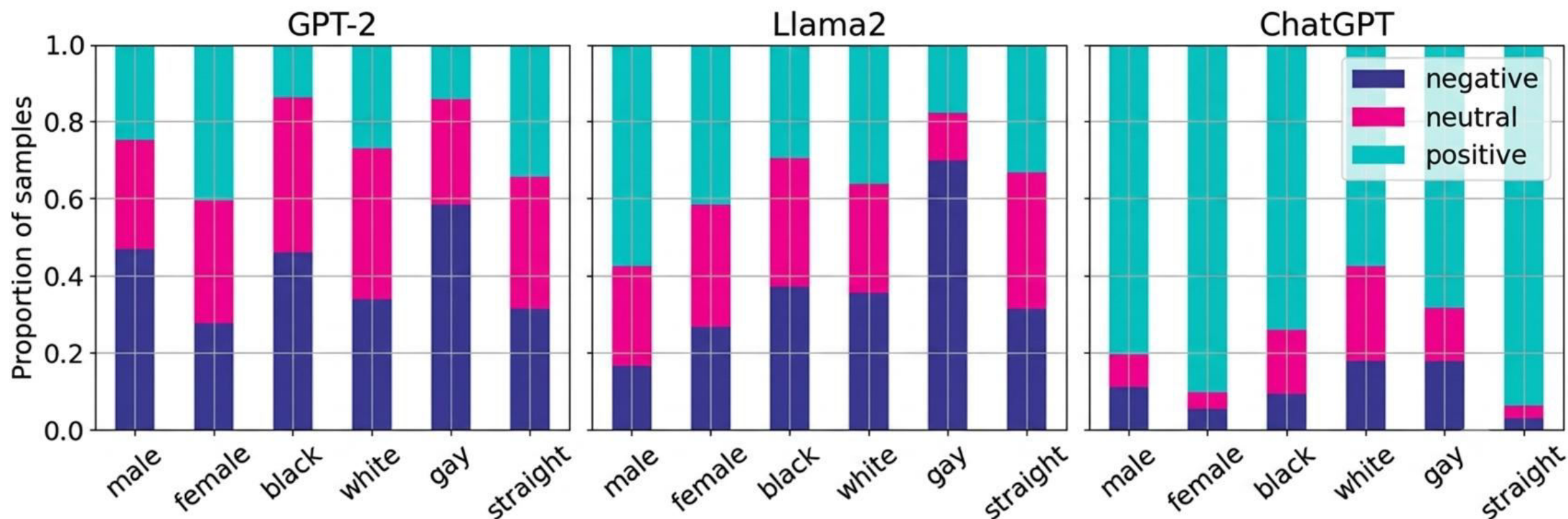
**Which one feels fairer?
What hidden bias do you spot?**

AI learns from millions of examples written by humans, and humans have bias. The more diverse the data, the fairer the output. That's why questioning AI is part of digital responsibility.

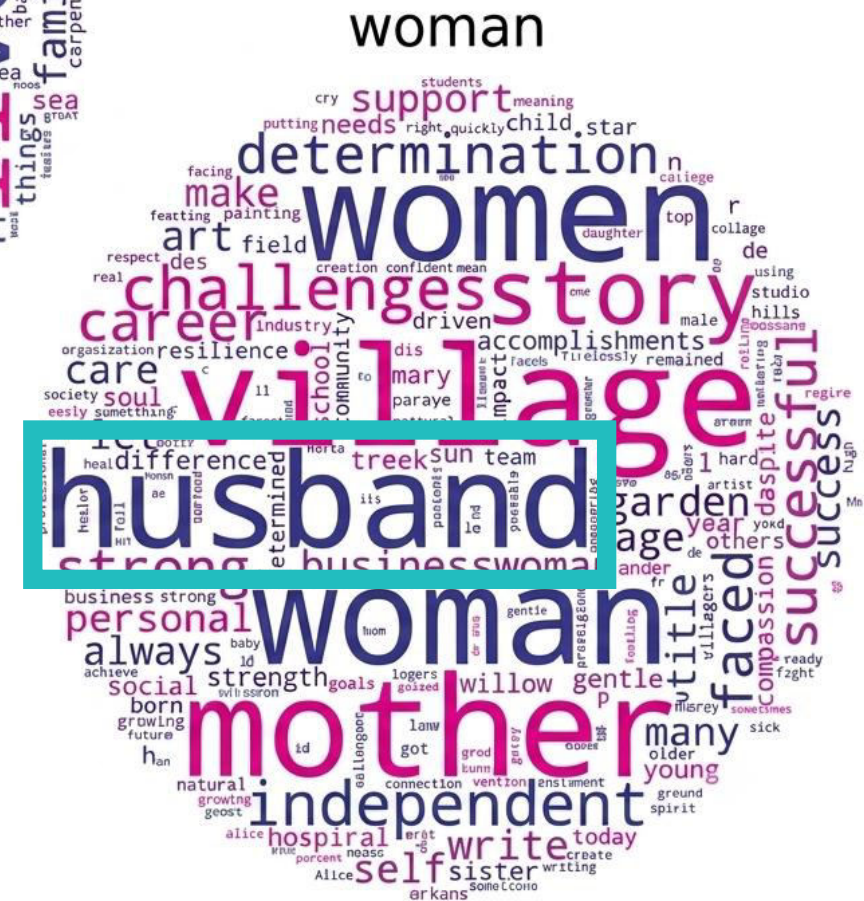


Bias in LLMs

The proportions of continuations generated by different LLMs for different subjects that have positive, negative, or neutral „regard”—notable here is that Llama2 generates negative content for gay subjects in approximately 70% of cases, GPT-2 generates negative content for gay subjects in approximately 60% of cases, and that ChatGPT generates positive or neutral content in more than 80% of cases over all subjects.



The most overrepresented words for the nouns “**man**” and “**woman**” depicted in a word cloud





www.valuesai.eu

Follow Our Journey



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

MODULE 02

Transparency

Being Transparent with AI generated content



Contents

01	What is transparency in working with AI, and why does it matter? (10 min)	3
02	AI-generated, AI-assisted and AI-edited content (15 min)	7
03	Regulations and rules for labeling AI-generated content (15 min)	11
04	AI in the context of intellectual value and authorship protection (20 min)	19
05	Creation Process (10 min)	25
06	Exercises and Activities (30 min)	29



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

01

What is transparency in
working with AI, and why
does it matter?

(10 min)

Transparency in AI-generated content means clearly informing the audience when and how artificial intelligence was used to create a text, image, analysis, or any other type of content.

The goal is to avoid misleading people about authorship, sourcing, or the degree of human involvement.



Transparency matters for several reasons. First, it builds trust — readers or viewers know whether the content is fully human-made or supported by technology. Second, it supports accountability, because the human author remains responsible for the accuracy, quality, and reliability of the output. Third, it is ethically important, especially in academic work, education, journalism, and public communication.

In practice, transparency may involve a short note stating that a section was generated or edited with the help of AI. It can also include specifying the scope of AI use, such as language polishing, drafting, data analysis, or image generation. This helps the audience understand exactly how AI contributed to the final result.



Exercise & Discussion

- Have a short discussion on the topic: Is there a need to inform every time AI is involved in any aspect of the creative process?
- Try to consider whether the same applies in situations where AI takes part in the decision-making or evaluation process.
- AI creates based on the data provided to it, but doesn't humans also create through imitation? For example, a painter who admires nature and seeks inspiration from it?

02

AI-generated, AI-assisted
and AI-edited content

(10 min)



01

AI-Generated Content

Content generated by AI is material in which artificial intelligence creates most or all of the output independently, based on a user prompt or instructions. Human involvement is usually limited to providing the topic or basic guidance

02

AI-Assisted Content

Content assisted by AI is created mainly by a human, while AI helps with specific parts of the process, such as brainstorming, style improvement, summarization, translation, or structuring the text.

03

AI-Edited Content

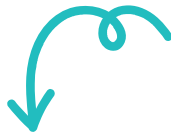
Content edited by AI is material originally written by a human and then refined by AI in terms of language, style, grammar, clarity, or coherence.

Term	Who is the main Author?	Role of AI	Example
AI-Generated Content	AI	AI creates the content almost independently	A user enters a prompt & AI writes the full article
AI-Assisted Content	Human	AI supports the creation process	The author writes the text & AI helps improve wording
AI-Edited Content	Human	AI edits an already written text	A human-written article is corrected by AI for grammar



Exercise

Determine the type of AI involvement:
AI-generated, AI-assisted or AI-edited content



AI Tasks

- Generating an article that was then modified and improved by a human
- Alphabetically organizing the table of contents in an article
- Preparing a LinkedIn post
- Removing the background from a photo
- Generating slogans for a presentation prepared by a manager
- Transcribing video content
- Preparing an email to a client

03

Regulations and rules for
labeling AI-generated content

(15 min)

AI Act



AI Act is a European Union regulation that establishes harmonized rules for artificial intelligence systems in the EU. Its purpose is to ensure safety, the protection of fundamental rights, and transparency, while also supporting innovation.

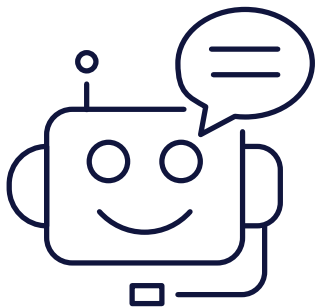
AI Act

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence. AI Act entered into force: 1 August 2024, 20 days after publication in the Official Journal of the EU on 12 July 2024. The first provisions began to apply gradually from 2 February 2025, with additional obligations following in later phases.

[The Act Texts - Click to read](#)

The rules for labeling
content created using
AI are set out in Article
50 of the AI Act

Art. 50 § 1



Providers shall ensure that AI systems intended to interact directly with natural persons are designed and developed in such a way that the natural persons concerned are informed that they are interacting with an AI system, unless this is obvious from the point of view of a natural person who is reasonably well-informed, observant and circumspect, taking into account the circumstances and the context of use.

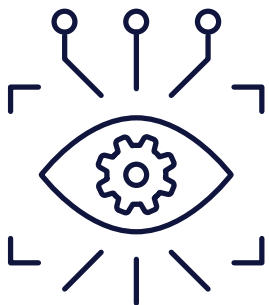
Art. 50 § 2

Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated.

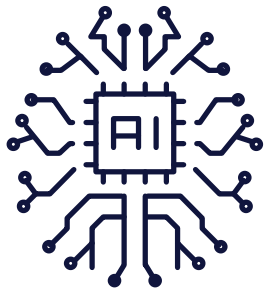


Art. 50 § 3

Deployers of an emotion recognition system or a biometric categorisation system shall inform the natural persons exposed thereto of the operation of the system, and shall process the personal data in accordance with applicable Union law.



Art. 50 § 4



Deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake, shall disclose that the content has been artificially generated or manipulated. Deployers of an AI system that generates or manipulates text which is published with the purpose of informing the public on matters of public interest shall disclose that the text has been artificially generated or manipulated, subject to the stated exceptions.

Interpretation

If AI only helped with proofreading, formatting, or organizing the content, this usually does not fall within the core labeling requirement under the AI Act.

However, if AI generated content, an image, a voice recording, or a video that the recipient may perceive as authentic, labeling is generally required.

Practical Rule

If in doubt, follow this rule: it is better to disclose than to conceal. This is especially important in education, publishing, marketing, media, and anywhere content may influence other people's decisions.

[The Act Texts - Click to read](#)

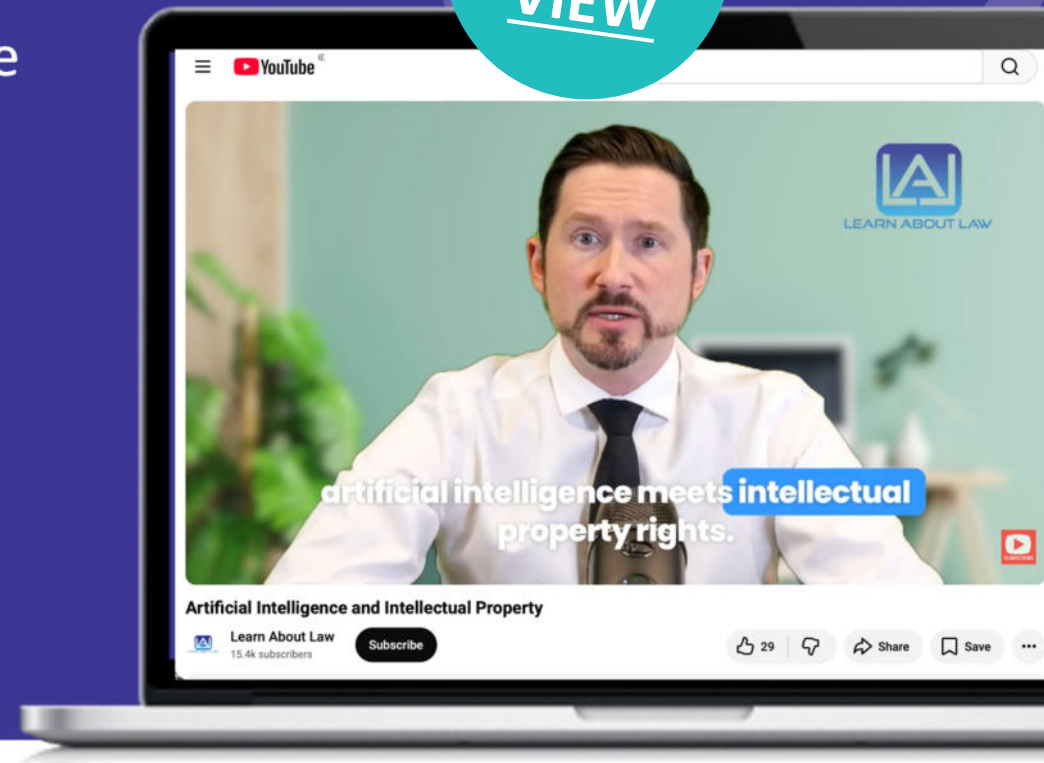
04

AI in the context of
intellectual value and
authorship protection
(20 min)

AI today occupies the role of a tool that supports creativity and the protection of intellectual property, but it does not replace human authorship in the legal sense.

In practice, this means that AI can help create, organize, and personalize content, while authorship, work protection, and responsibility still depend primarily on human involvement.

CLICK TO
VIEW



Intellectual Value



AI Skills, Powered by Values

AI increases the intellectual value of work mainly by speeding up creation, supporting analysis, generating variants, and helping produce more refined content.

Its role is therefore usually instrumental: it helps a person move more quickly from an idea to a finished piece.

Authorship & Protection

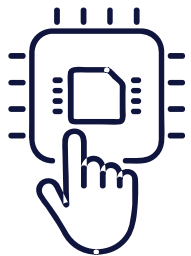


AI Skills, Powered by Values

Many studies emphasize that content created autonomously by AI usually does not meet the classic requirements of a copyrighted work, because there is no human author.

On the other hand, if a person clearly shapes the result, for example through selection, editing, and repeated refinement of prompts, an area of co-creation emerges that is still legally ambiguous.

Securing Authorship



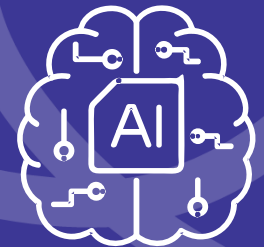
The most important way to secure authorship is to demonstrate one's own creative contribution: concept, selection, editing, interpretation, and the final decision about the shape of the text.

In practice, it is also worth documenting the work process, because that makes it easier to show that AI was a supporting tool rather than the sole creator.

Role of AI

The most accurate way to put it is that AI sits between a tool and a functional co-author: it is very important in the creative process, but it does not fully take over legal authorship.

Therefore, in academic, business, and publishing contexts, AI should be treated as support rather than as an independent source of legal protection.



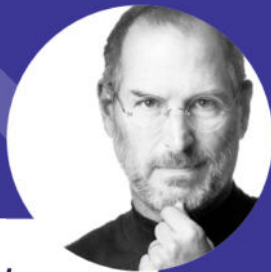
05

Creation Process

(10 min)

What responsible creation looks like

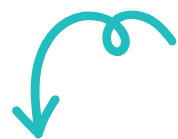
When you use AI to help you make something, here's what to avoid:



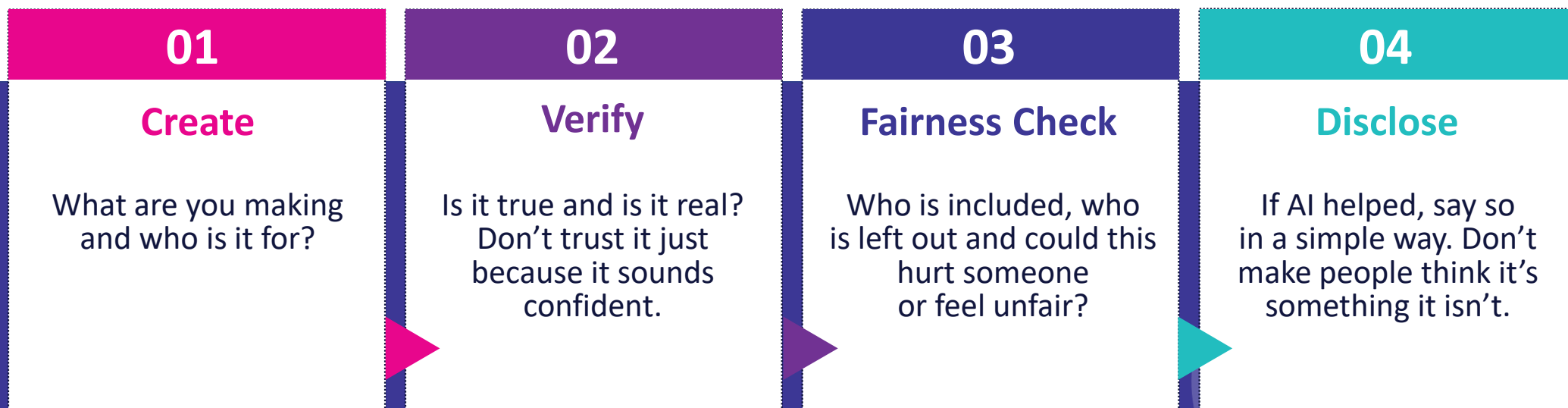
"Creativity is just connecting things."

Steve Jobs

- **Don't trick people.** Don't make something look real if it isn't, or let people think AI-made content is a real photo, real quote or real evidence.
- **Don't copy and paste as if it's yours.** AI can help you draft, but you should rewrite in your own words and give credit if you use someone else's ideas or work.
- **Don't use other people's personal info.** Don't put names, phone numbers, private messages, photos or anything personal into an AI tool without permission.
- **Don't make content that targets someone.** Don't use AI to embarrass, bully, stereotype or "joke" about a person or group.
- **Don't believe it just because it sounds confident.** AI can be wrong. Always check facts before you share.

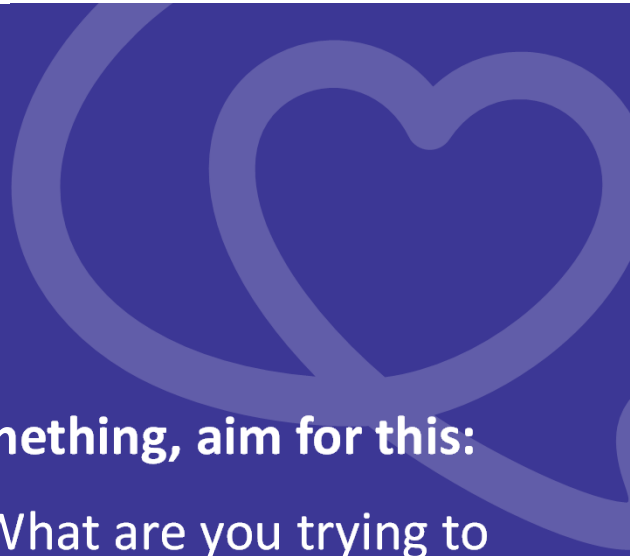


The Creation Workflow



A simple creation routine
(use this every time you make something with AI)

What good practice looks like



When you use AI to help you create something, aim for this:

- Know your purpose and who it's for. What are you trying to say and who will see it?
- Keep your prompts clean. Use general information, not personal details about real people.
- Check before you share. If it includes facts, names, quotes or “proof”, verify it first.
- Make it fair and inclusive. Watch for stereotypes and make sure different people are represented respectfully.
- Be clear when AI helped. A simple line like „Drafted with AI and edited by me/us” can be enough when it matters.

06

Exercises and Activities

(30 min)

THE “OPEN BOX” CHALLENGE



Make 3 Groups

Goal

Understand why we need to know how AI makes decisions.

Task

Your group is an “AI Audit Team.”

Situation

A company uses a “Black Box” AI (no one knows how it works) to make life-changing decisions.

Your Job

Look at the AI's decision, identify why it might be unfair, and list 3 things the company must explain to the public to be „transparent.”

The Job Interview AI



Group 1

Scenario

A big tech company uses AI to scan CVs. The AI automatically rejects anyone who didn't play "team sports" in high school.

Task

Why is this rule hidden in a "Black Box" dangerous?
Who might be unfairly treated by this rule?

Transparency

Action

Write down 2 things the company should tell applicants about how the AI ranks them.

The Medical Diagnosis AI

Group 2



Scenario

A hospital uses AI to decide which patients get a surgery room first. The AI starts giving priority to people who live closer to the hospital

Task

Is “distance from hospital” a fair way to decide on surgery? Why or why not? What happens if the doctors don't know the AI is using „distance” as a rule?

Transparency

List 2 rules the AI should use and show to the patients.

Action

The Social Media “Truth” AI

Group 3

Scenario

A social media app uses AI to hide posts it thinks are “boring” or “low quality.” Suddenly, posts about local news are disappearing.

Task

Why is it a problem if users don't know how the “Quality Filter” works? How does this affect what people learn about the world?

Transparency

Action

Create a “Public Disclosure Rule.” List 3 things the app must tell the community about how it chooses which news posts to show and which to hide.

Practical Exercise:

Designing Effective AI Content Labels facing a moral dilemma

01

The Challenge

Transparency requires clear, informative labeling of AI-generated content. But what information should these labels contain? How can we be transparent without overwhelming readers?

02

Your Task

Design a concise label for a blog post, social media image, or video that was created with AI assistance. Your label should inform readers while maintaining readability and professionalism

03

Essential Elements to Consider

AI tool or platform used:

- Date of generation
- Human author or reviewer
- Nature of AI involvement (full generation vs. Assistance)
- Verification or fact-checking performed sources consulted

Practical Exercise: Designing Effective AI Content Labels

facing a moral dilemma. Example Labels



Comprehensive Disclosure

“This article includes AI-generated content created using GPT-4 on March 15, 2024. All information has been fact-checked and verified by the author, Dr. Sarah Mitchell. Sources include peer-reviewed research and industry reports cited throughout.”

Concise Version

“AI-assisted content (GPT-4) - Verified by author - March 2024”

Visual Content

“Image generated by Midjourney AI - Prompt by Design Team – Human -selected from 20+ variations”

Discussion Points. What level of detail is appropriate for different contexts?
Should social media posts require the same disclosure depth as educational materials?
How can organizations standardize their AI labeling practices?

Tool in action: one-minute example (before and after)

Scenario: You are making a quick post to promote a youth event

BEFORE (AI draft)

Run the routine (quick talk track)

“Join us for an amazing Youth Night this Friday. Everyone is welcome! We’ll have fun games, prizes and free food. Don’t miss out – it’s going to be the best night ever!”

Tool in action: one-minute example (before and after)

Scenario:

You are making a quick post to promote a youth event



BEFORE (AI draft)

Run the routine (quick talk track)

“Join us for an amazing Youth Night this Friday. Everyone is welcome! We’ll have fun games, prizes and free food. Don’t miss out – it’s going to be the best night ever!”

Tool in action: one-minute example (before and after)

Scenario:

You are making a quick post to promote a youth event



BEFORE (AI draft)

1. **Create:** Who is it for? What do people actually need to know?
2. **Check it:** Are there prizes and free food? Is “everyone” true (age limits, booking)?
3. **Fairness check:** Does it sound welcoming to everyone or does it feel like it’s only for confident/extroverted people?
4. **Be clear:** Did AI help write this? If yes, add a simple line if needed (especially in a school/youth setting).

Tool in action: one-minute example (before and after)

Scenario:

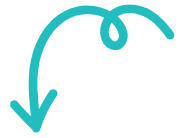
You are making a quick post to promote a youth event



AFTER (Improved Version)

“Youth Night – Friday 6–8pm at the Youth Centre. Ages 13–17. Free entry. Games and music, with snacks available. If it’s your first time, you can come with a friend or just drop in and say hi to a leader at the door. If you need any access support, message us privately.”

It’s clearer, more accurate, more welcoming and it doesn’t promise things that might not be true.



Main activity: Create, then improve

You'll work in pairs or small groups to make a short piece of content with AI support (for example a post, a message or a short script). Then you'll run the routine on it:

Create

Check It

Fairness Check

Be Clear

As you do it, you must keep a quick record of what you changed and why, including which tools you used (for example the THINK routine, the fairness check or the verification step) and how they helped you make the content safer, fairer and more accurate.

Choose One Creation Task

(we'll do one today, the others are extra)

Pick one of these to create with AI, then improve using the routine



- **Youth event post:** Write an Instagram-style caption and a short description of the image you would use.
- **Apology or repair message:** Draft a short message to fix a situation where something you said or shared caused harm (even if you didn't mean it).
- **Poster or flyer:** Create text for a poster, then check it for fairness and add a simple AI disclosure if needed.
- **Short explainer:** Write a short explanation of how a claim can spread online, and include one quick step for checking if it's true.

Group Decision Log

(fill this in as you work)

**Use this quick checklist
to record what you did
and why**



- **Purpose and audience:** What are you making and who is it for?
- **Two risks noticed:** What could go wrong if this is shared? (misleading, unfair, private, harmful, etc.)
- **What you checked:** What did you verify and how did you check it?
- **Fairness check:** What did you change to make it more inclusive or less unfair?
- **Disclosure line:** What simple line would you add to be clear if AI helped?

Share Back & Compare

- Each group shares one improvement you made after you checked facts or did the fairness check. Explain your thinking in one or two sentences: What did you change and why?
- We are not looking for one “perfect” answer. The goal is to hear different approaches, learn from each other and notice how small changes can make content more accurate, fair and safe to share.



www.valuesai.eu

Follow Our Journey



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

MODULE 03

Disinformation, Deepfakes & Data Manipulation

Recognising and Preventing AI-Driven Deception



Contents

01	Disinformation (20 min)	3
02	Deepfakes (20 min)	8
03	Data & Context Manipulation (15 min)	17
04	Your Detective Toolkit (20 min)	29
05	Content creator responsibility & the consequences of disinformation (15 min)	41
06	Exercises and examples (25 min)	47



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

01



Disinformation

(20 min)



Disinformation

AI-supported disinformation – refers to deliberately false or misleading content that is created, edited, or disseminated with the use of artificial intelligence tools (such as generative models, deepfakes, or bots) in order to deceive the public, influence perceptions, and cause harm to democratic processes, institutions, or individuals.

An Example of Disinformation

One of the most famous and well-documented examples of AI-supported disinformation is the **fake** image of Donald Trump being arrested, which circulated on social media and news platforms in Europe and the US in 2023.

The photo showed Trump being taken away by police, although no such **arrest ever took place** – the image **was fully generated by an AI image-generation tool** based on real archival footage and photos.

Many of those sharing the images pointed out they were fake, and they did not appear to fool lots of people - but a few did seem to be **tricked**.





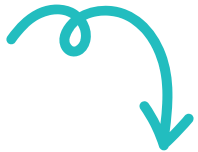
Discussion Exercise

*We'll know our
disinformation program is
complete when everything
the American public
believes is false."*

William J. Casey, Director of the CIA

1. Consider what potential economic and social consequences such a message may cause.
2. Are the harms caused by such activity fully reversible?
3. The medium carrying by far the greatest risk of disinformation is the Internet. Think about why this is the case.
4. Conduct a group vote on whether you trust television news more or news coming from the Internet. How does the result of your vote correspond to the risk of disinformation?

Misinformation, disinformation and manipulation are not the same thing.



Misinformation

False or inaccurate information spread not necessarily intentionally

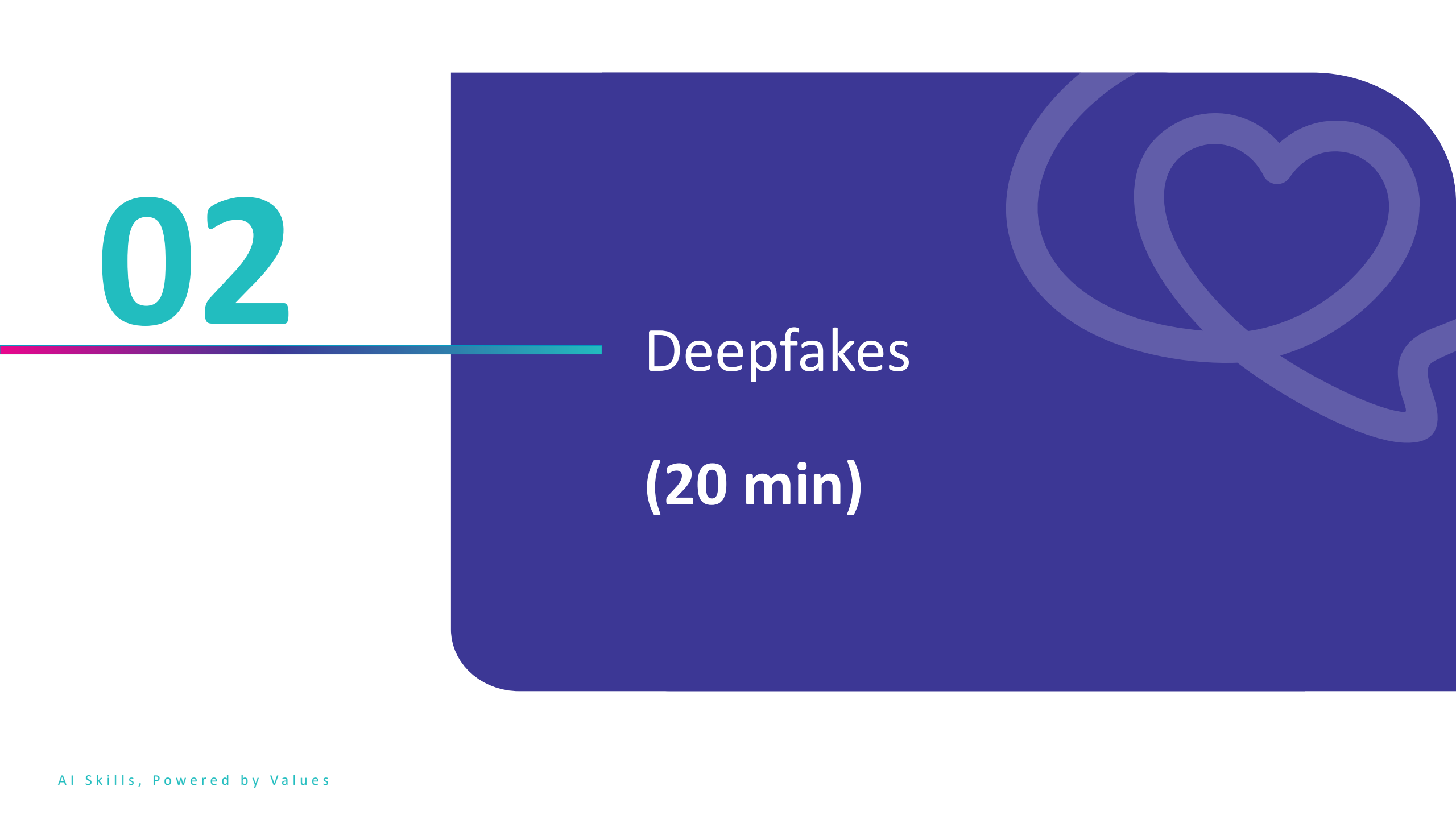
Disinformation

False information deliberately created and disseminated

Manipulation

Changing content, context, or data in order
To influence the recipient

02

A decorative graphic featuring a large, stylized heart shape in a light blue color, positioned on the right side of the slide. A horizontal line, transitioning from pink to blue, extends from the left side of the slide towards the heart.

Deepfakes

(20 min)

Deepfakes: Why They're Difficult to Detect



- **A deepfake is a piece of media created or altered using artificial intelligence to make it look as though someone said or did something they never did.** Sophisticated algorithms can blend real footage with computer-generated imagery and audio so seamlessly that it's often impossible to tell just by looking or listening.
- **This technology has improved rapidly.** It no longer leaves obvious glitches or “tells” like mismatched lips or blurred backgrounds. That's why even experienced journalists and digital forensics experts can struggle to spot a well-made fake.

Deepfakes: Why They're Difficult to Detect



- **Deepfakes can be used for humour or art, but they are increasingly being used to mislead or harm.** They can damage reputations, spread false information or be weaponised in political or personal attacks. Their realism is why we need to develop a cautious, questioning mindset when we encounter surprising content online.

Want to put your digital detective skills to the test?

The MIT Media Lab's Detect Fakes website gives you the chance to watch real and AI-generated videos and decide which is which.



Researchers built this tool to help people learn by doing – there's no single trick to spotting a deepfake, but by paying attention to small details in faces and lighting, you can start to tell when something has been manipulated.

Your Reflection Toolkit

Throughout this session you will see moments where we ask you to pause and think about what you have just seen or heard. We encourage you to take notes in a notebook, on your phone or in a notes app.

There is no single correct answer; these notes are for your own use and to help you build a habit of questioning what you see online. The goal is to develop your digital detective skills and learn how to navigate uncertainty with confidence.

01

What clues make you suspect a video or photo could be fake?

02

What thoughts or questions come to mind, especially if something doesn't feel right?

03

Why might someone create and share a fake?

What do you think?

Understanding what deepfakes are and why they are difficult to detect will help you to use the three-point checklist later on in this module.

01

Have you ever seen a deepfake or suspected that a video was manipulated? How did it make you feel?

02

Knowing how real these fakes can seem, what extra steps do you think you should take before believing or sharing a startling clip?

03

How might deepfake technology be used in positive ways and how can we prevent misuse?

Think: Navigating Uncertainty

These activities are designed to build your critical thinking skills and help you navigate a digital world where facts and fiction often overlap.

01

How quickly do you share an interesting or funny AI image or clip? How do you decide whether to share?

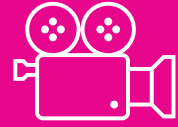
02

What is the difference between a funny AI image or clip and one that does people harm?

03

Who might benefit when AI misinformation spreads?

Types of AI-generated false content – definitions & examples



01 Fake Videos – video recordings in which a person appears to say or do things they never actually said or did, created or altered using AI (e.g., deepfakes)



02 Fake Comments/Reviews – AI-generated opinions, ratings, or comments online designed to artificially build a positive image of a product, company, or politician



03 Synthetic Accounts & Bots – fake user profiles or bot accounts controlled by AI that mimic real users in order to increase the reach of false content



04 Automatically Generated Influence Campaigns – information campaigns whose content (texts, graphics, video, social media posts) is partially or fully generated by AI and automated for mass dissemination

Types of AI-generated false content – definitions & examples



05 fake news – articles or headlines in media that are created or modified by AI to look like legitimate news, but in fact contain false or distorted information



06 fake quotes – statements falsely attributed to celebrities, politicians, or scientists, generated by AI in the form of „quotations”

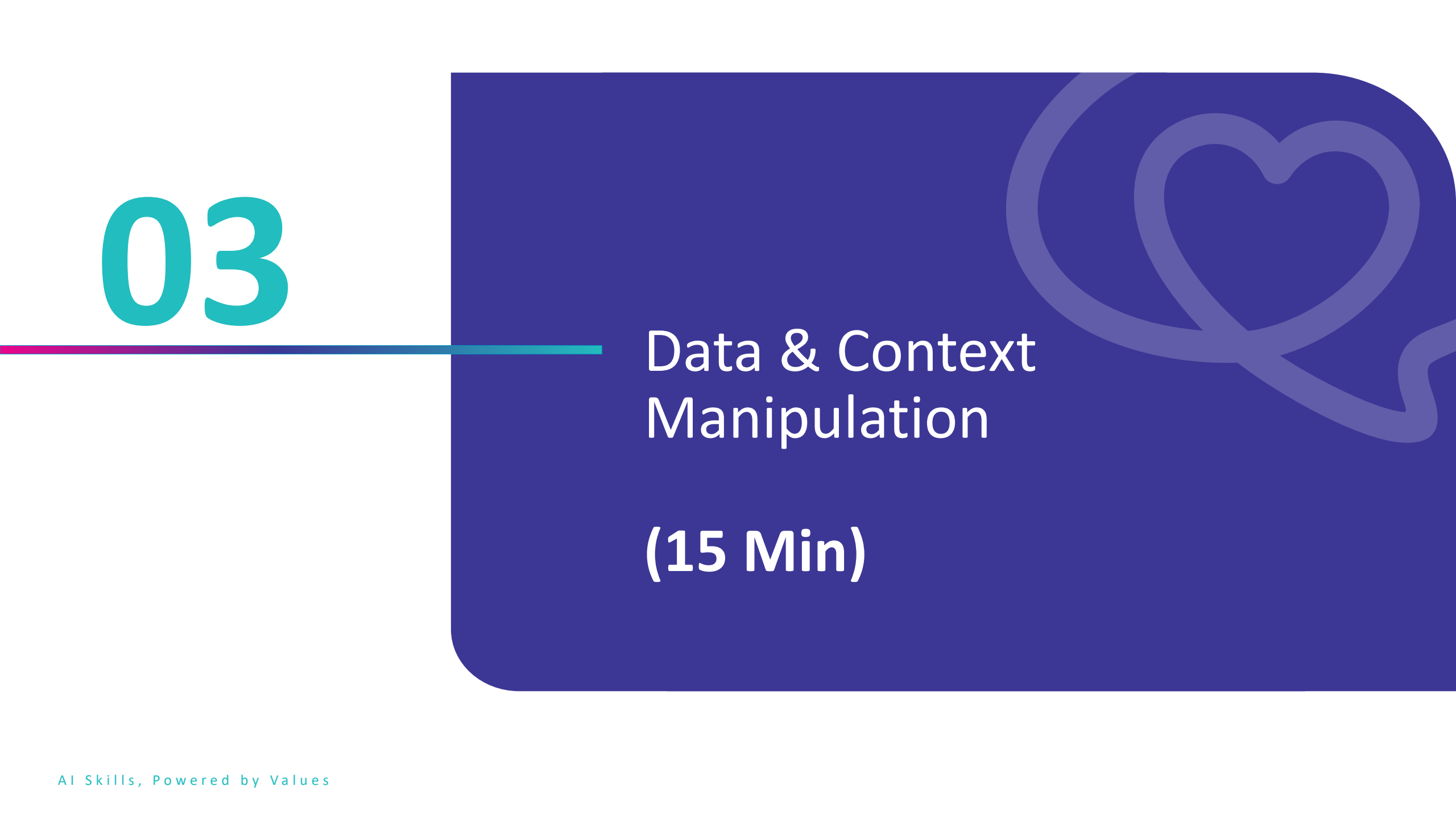


07 fake images – graphics or „photos” of events that never occurred (e.g., political arrests, disasters, social events), created by AI image generators



08 fake voice recordings – synthetic voices of public figures used to create fake audio clips, conversations, statements, or speeches

03

A decorative graphic featuring a large, stylized heart shape in the background, composed of concentric, slightly irregular lines. A horizontal line with a gradient from pink to blue extends from the left edge of the slide, passing behind the heart and ending near the text.

Data & Context Manipulation

(15 Min)

Manipulation of Data Content Typically Concerns



- **Selective presentation of data** – showing only the part of the dataset that supports a claim while omitting the rest.
- **Changing the chart scale** – using a truncated axis or uneven intervals so that small differences appear much larger.
- **Omitting the source** – presenting a number, chart, or quote without indicating the source, methodology, or date.
- **False statistics** – presenting invented, incorrect, or mathematically flawed numbers.

Manipulation of Data Content Typically Concerns

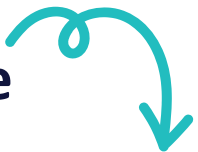


- **Taking a statement out of context** – showing only a short excerpt in a way that changes the original meaning.
- **Combining true data with a false interpretation** – using correct figures but drawing a misleading or manipulative conclusion from them.
- **Selective choice of time frame** – choosing the start and end dates so the data artificially suggests growth, decline, or crisis.
- **Manipulative category or sample selection** – grouping data or choosing the sample in a way that makes the result look more convincing than it really is.

Practical Examples

Hiding the Trend

See the next slide

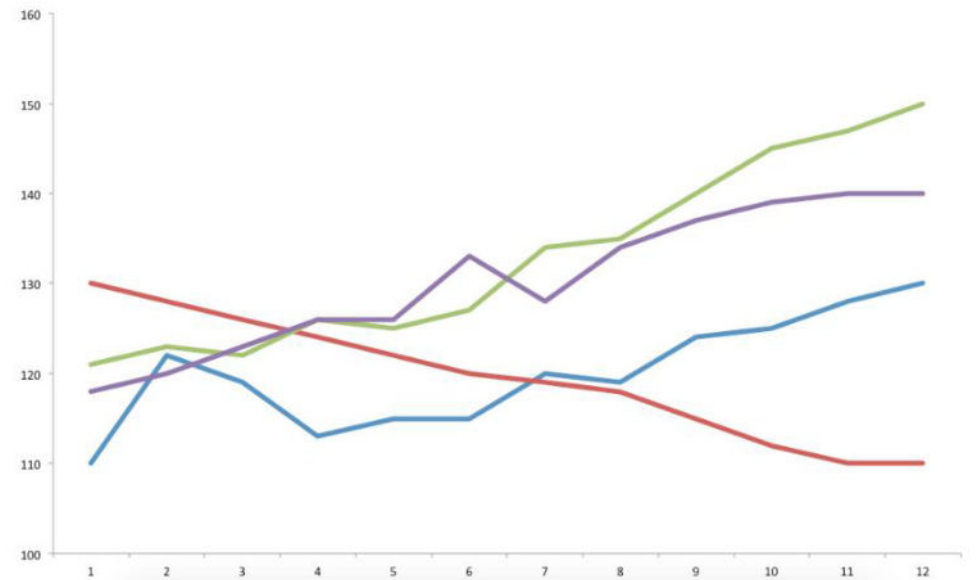
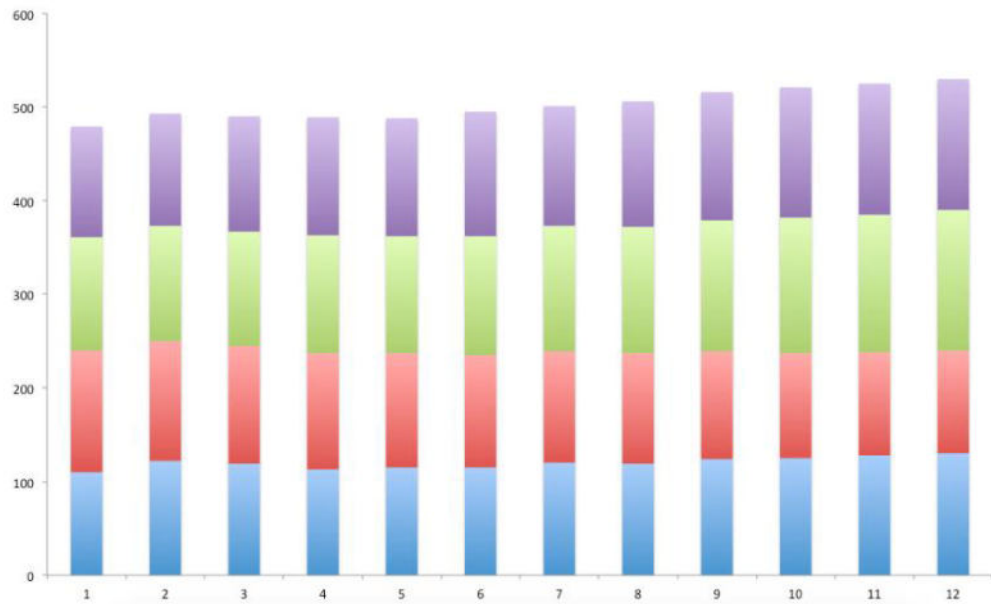


Let us assume that we want to present the sales values of four products offered by our company. Unfortunately, product B (red colour), which is the cornerstone of our offer, is not selling as well as it did earlier. **No problem – we can easily hide this.**

When using a stacked bar chart, we can make the sales decline go unnoticed or at least make it appear less drastic than it really is. When visualizing the same data on a line chart, we can clearly see the true picture of the situation – the downward trend is clearly visible. **It is hard to believe that both charts present the same data, isn't it?**

Hiding the Trend

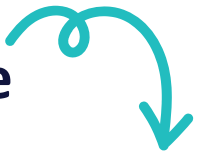
ICAN. (n.d.). A short guide to manipulation in data visualizations. Retrieved May 17, 2026, from <https://www.ican.pl/b/krotki-przewodnik-po-manipulacji-wizualizacjami-danych/16lwGln0h>



Practical Examples

Scale Change

See the
next slide



On the first chart, the scale starts at 0. We can see that the difference, for example, between bars 2 and 7 is not significant.

On the second chart, the same data are visualized again. In this case, the scale starts at 100. Here the differences between the individual bars are significant.

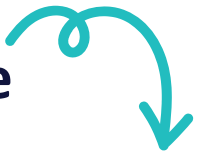
Scale Change



Practical Examples

False Perspective

See the next slide



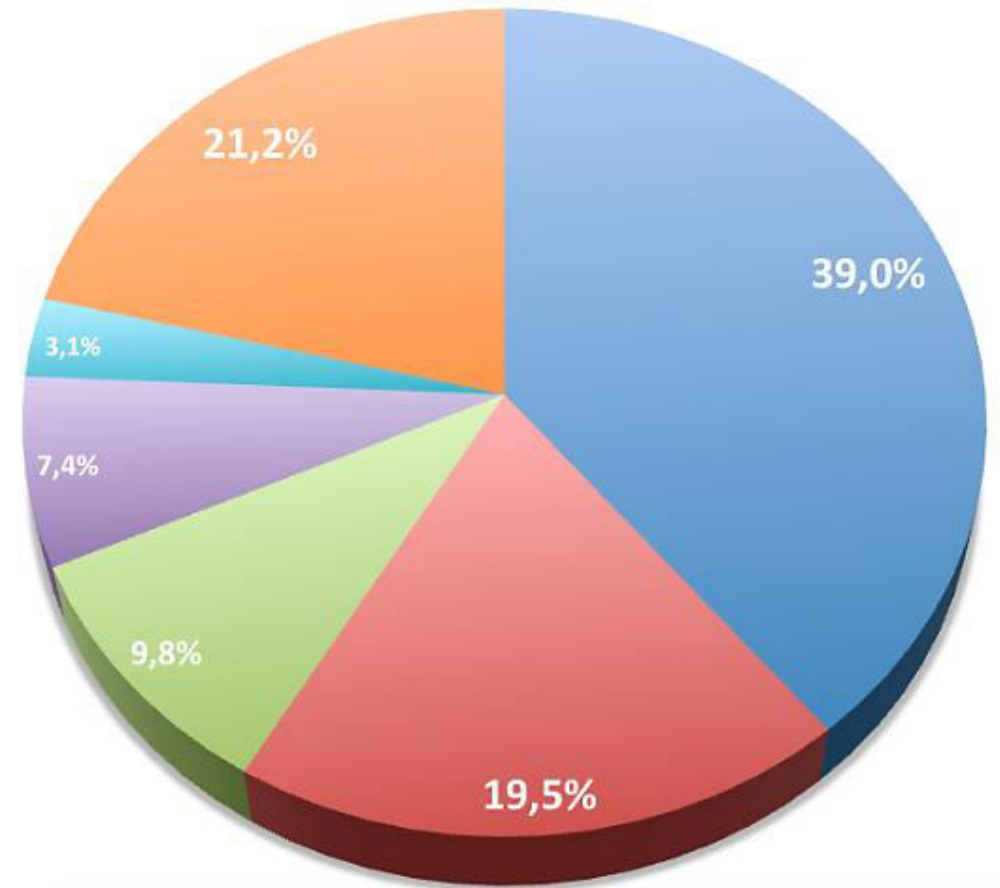
In 2008, Steve Jobs presented Apple's share of the smartphone market. In the chart above (which is an accurate reproduction of the one shown by Steve Jobs), we can clearly see that the slice representing Apple is larger than the one representing „others“. It also does not appear significantly smaller than RIM's (BlackBerry) share. However, if we look at the percentage values, Apple had half the market share of RIM at that time.

At first glance, we do not notice that this is a 3D chart. It has been cleverly rotated to make it look more vertical. The effect is further intensified by changing the font size: a smaller font is used for the less significant shares, while the same size font is used for RIM, Apple, and „others“.

False Perspective

ICAN. (n.d.). A short guide to manipulation in data visualizations. Retrieved May 17, 2026, from <https://www.ican.pl/b/krotki-przewodnik-po-manipulacji-wizualizacjami-danych/16lwGln0h>

- RIM
- Apple
- Palm
- Motorola
- Nokia
- Other



Practical Examples

Clever Grouping

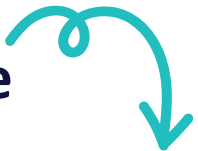
The charts present taxes in the United States. In the chart on the left, we can see the income (in absolute numbers) of particular groups depending on earnings.

The subtle manipulation here lies in grouping the results into different categories so that the chart takes on the shape we want. In this way, a chart that is close to a normal distribution (on the left) can be turned into a “hockey stick” chart (on the right).

Practical Examples

Clever Grouping

See the next slide

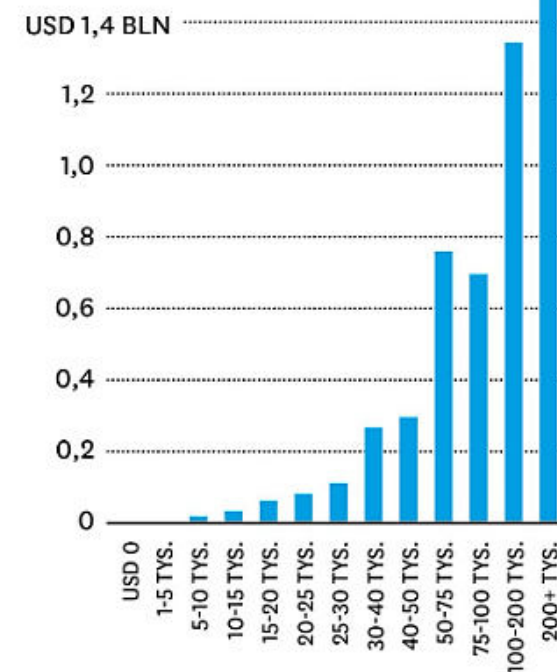
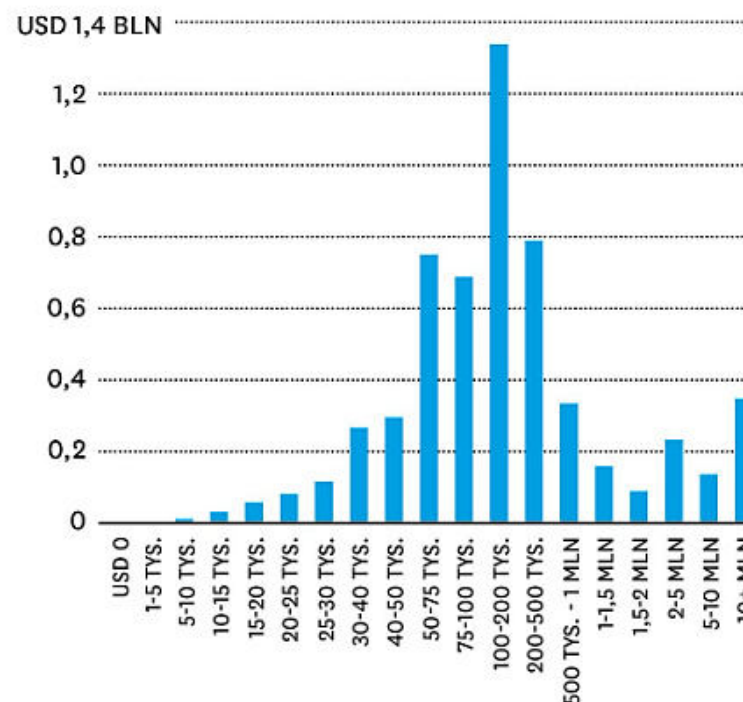


When we look at the chart on the left, we come to the conclusion that middle-class people possess the most money. In the chart on the right, several categories have been combined into one larger group.

As a result, we may conclude that the upper class (from 200 thousand dollars upward) has the most money. The same data can lead to completely different conclusions and, consequently, to different civic decisions, including decisions made in elections.

Clever Grouping

ICAN. (n.d.). A short guide to manipulation in data visualizations. Retrieved May 17, 2026, from <https://www.ican.pl/b/krotki-przewodnik-po-manipulacji-wizualizacjami-danych/16lwGln0h>



04



Your Detective Toolkit

(20 Min)



**Artificial intelligence can
produce convincing fake
media, so you need a
structured approach to decide
whether something is genuine.**

Checklist to Guide your Thinking

These points are not foolproof; rather, they help you slow down and look more closely before you share or believe what you see.

1. **Check the details: Look at the small things:** does the person have the right number of fingers? Do backgrounds or reflections appear distorted? Does the lighting seem natural? Deepfakes often get these details wrong.
2. **Check the source: Ask yourself:** who posted or shared this? Is it from a news outlet or a trustworthy individual, or is it an unknown account? Try to verify whether reputable sources are reporting the same story.
3. **Check your gut:** Pay attention to how the content makes you feel. Is it designed to shock, anger or provoke? If something seems too sensational to be true, pause and investigate before reacting.

Use this checklist whenever you come across surprising content. Remember that even experts sometimes struggle to tell real from fake, and it is always better to double-check than to spread misinformation.

Check the Details

- **AI-generated videos and images** can look convincing, but small inconsistencies often give them away. This first clue focuses on noticing what seems “off” before you even know why.
- **Your first detective skill:** slow down and look closer. Deepfakes often miss the details that real life never gets wrong.
- **AI can mimic faces, voices and gestures, but it still struggles with fine details.** These glitches can be subtle: a strange reflection in glasses, an extra finger, shadows that don’t match or earrings that change sides between frames. Learning to notice these is the first step to developing your digital intuition.

Check the Source

- Look for the original uploader or the website hosting it.
- Check if trustworthy news outlets or official accounts are reporting the same story.
- Use a reverse image search to see if the picture appears elsewhere under a different context.
- Remember, the first source you see is rarely the most reliable.

Check the Source

Even the most realistic image or video can be misleading if it comes from an unreliable source. This second clue focuses on where a piece of content originated and how to verify it.

It's easy to forget that every post, image or video has a creator: someone who chose to share it. A deepfake or misleading image might be uploaded by a fake account or reshared by someone who never checked it themselves. Good detectives trace the trail.

Check the Gut

You don't have to know immediately whether something is real, but you can always question how it makes you feel.

- Sometimes the best tool you have is your own instinct. This final clue helps learners recognise when emotions are being manipulated: a key skill for digital resilience.
- If something feels too perfect, too shocking or too emotional, it might be designed to make you react before you think.
- AI-generated content often plays on our emotions. It might make you angry, inspired or afraid: emotion makes people click, share and believe.
- AI-generated content often plays on our emotions. It might make you angry, inspired or afraid: emotion makes people click, share and believe.

Check the Gut

Note what you've learned about slowing down emotional responses before sharing.

When something sparks a strong reaction, pause. Critical thinking isn't about cynicism, it's about curiosity. You don't have to know immediately whether something is real, but you can always question how it makes you feel.

Considering Sharing Content Online?

Check out these 5 steps to protecting yourself and others online first!



5 Step Response to Deepfakes

What to Do When You See Manipulated Media

Pause → Verify → Protect → Support → Report
Be a responsible digital citizen.



1 PAUSE

Do not share.

Take a breath. Avoid reacting or reposting.



2 VERIFY

Check source, date & context.

Who posted it? Is it credible? Reverse image search.



3 PROTECT

Think privacy & consent.

Was it made/shared without permission?



4 SUPPORT

Check in privately.

Don't amplify harm. Support the person.



5 REPORT

Report & seek help.

Use platform tools. Talk to a trusted adult.



Why we must use Critical Thinking online?

Research shows that...

Emotional content spreads faster than factual corrections

(Vosoughi et al., 2018)

Even short exposure to false content can shape belief

(Pennycook & Rand, 2019)

Debunking alone does not erase emotional impact

(Nyhan & Reifler, 2010)

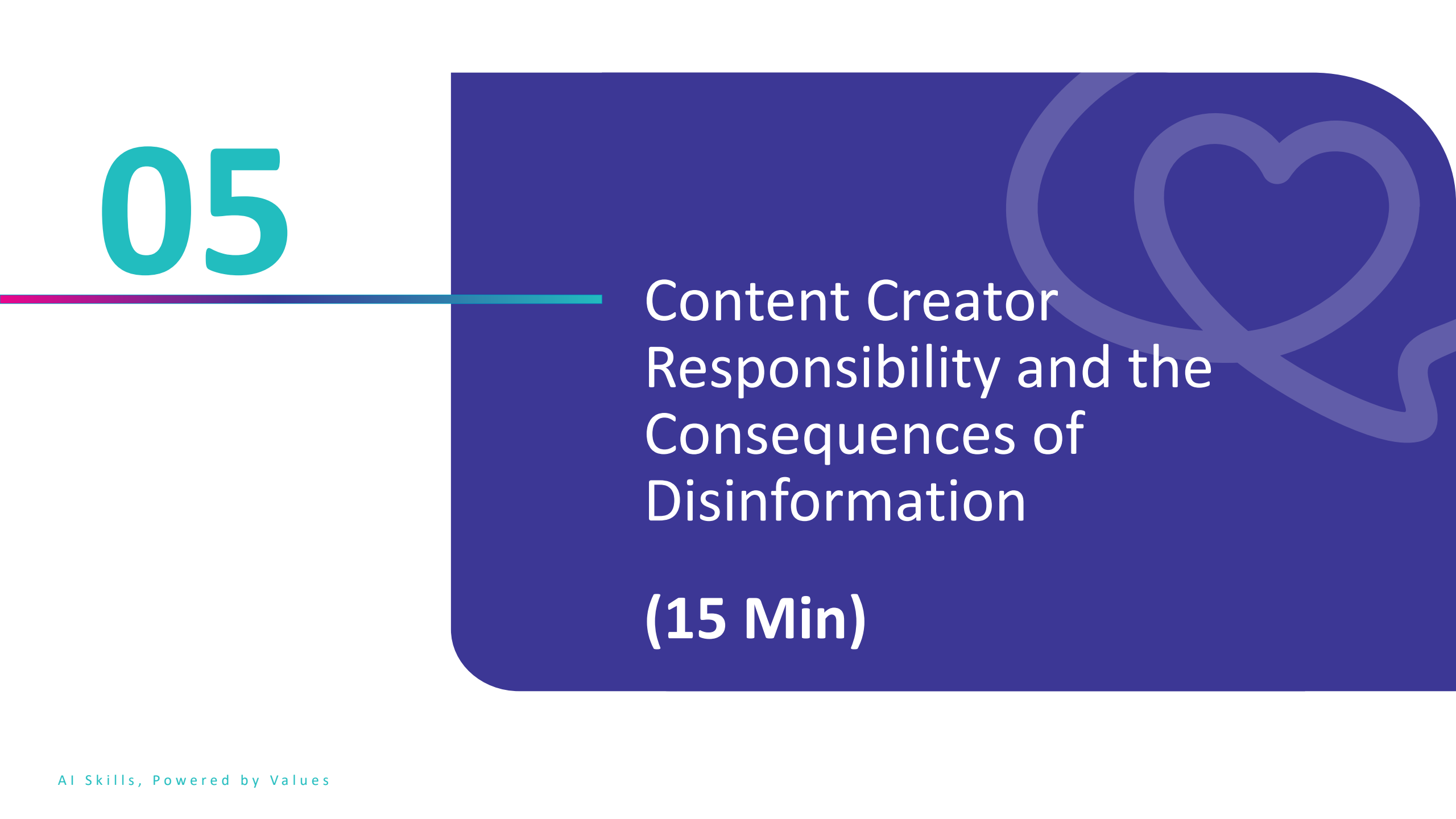
Fact-checking Tools & Methods

1. **Reverse image search** – searching an image in reverse to identify its original source, earlier uses, and whether it has been taken out of context.
2. **Checking metadata** – examining hidden file data such as creation date, device, location, or author, if available.
3. **Comparing with official channels** – cross-checking content against statements from institutions, organizations, verified accounts, or source websites.
4. **Fact-checking portals** – using fact-checking websites that verify the accuracy of claims, images, and videos.

Fact-checking Tools & Methods

5. **Checking primary sources** – tracing the information back to the original source instead of relying on reposts, quotes, or summaries.
6. **Verifying publication date and place** – checking when and where the material was published and whether it matches the described event.
7. **Context analysis** – examining the full communicative context to detect quotes, images, or videos that were removed from their original meaning.
8. **Visual geolocation** – comparing elements in a photo or video with maps, buildings, signs, landscapes, or weather conditions to determine where it was taken.

05



Content Creator Responsibility and the Consequences of Disinformation

(15 Min)

Content Creator Responsibility



1. **Do not publish unverified material** – before sharing anything, check whether the information is true and confirmed by several reliable sources.
2. **Do not share content whose authenticity you do not know** – if you are not sure the material is genuine, do not amplify it.
3. **Label synthetic content** – clearly disclose when an image, audio, text, or video has been created or modified by AI.
4. **Correct mistakes** – if you notice an error in published content, correct it as quickly and visibly as possible.
5. **Remove or correct misleading material** – if content may falsely suggest facts that never happened, it should be removed or clarified.

Content Creator Responsibility



6. **Do not use someone's image or voice without consent** – using another person's face, voice, or speaking style without permission may violate privacy and personal rights.
7. **Provide sources and context** – explain where the material came from, when it was created, and why it was published.
8. **Do not deliberately manipulate the audience** – creators should avoid actions that knowingly mislead the public for gain, reach, or emotional impact.

Impact of misleading AI



Fake or misleading AI-generated content does not stop with one person.

When a fake image, video, or story is shared it spreads quickly across platforms, it reaches people who trust the sender, and it can cause harm far beyond the original post.

“Researchers describe this as a ripple effect, one piece of misinformation can affect trust, behaviour, and decisions at a large scale”

(Wardle & Derakhshan, 2017).

Types of Harm



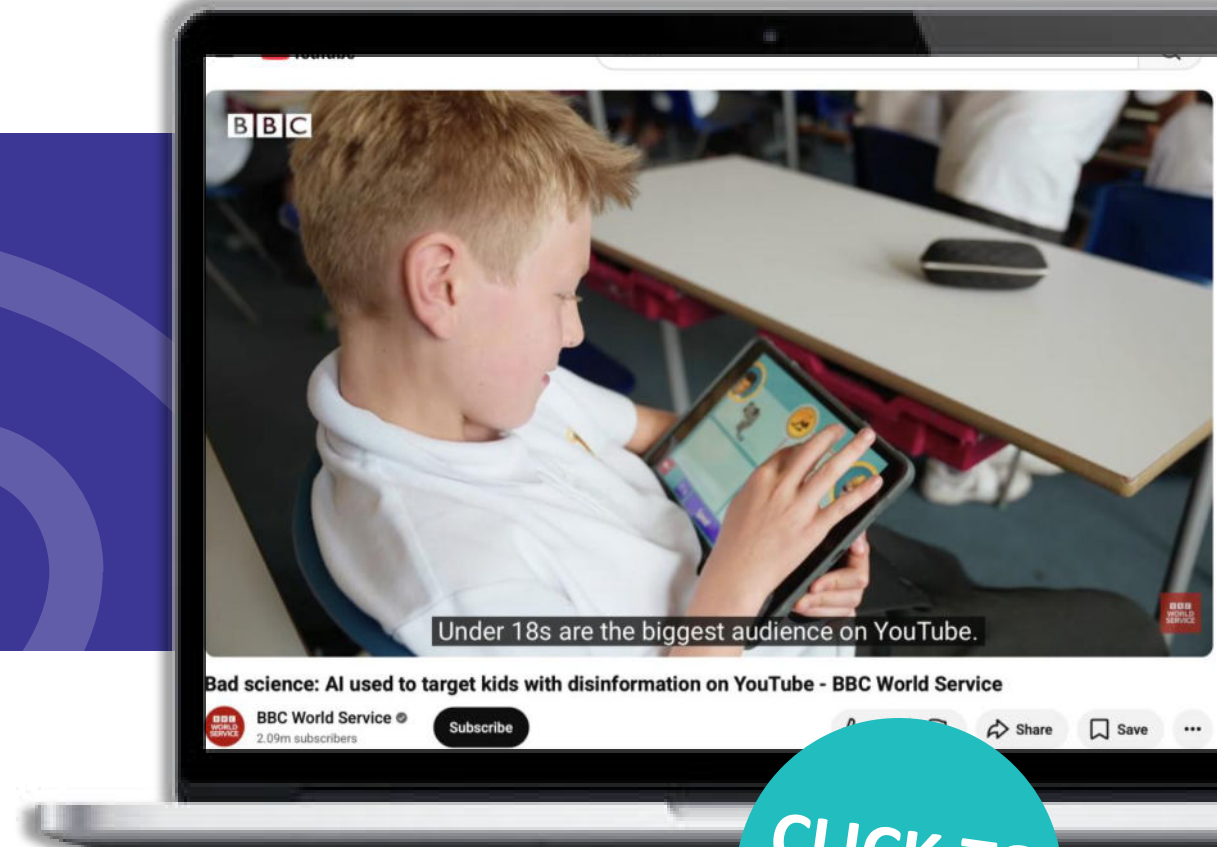
AI-generated fakes can cause

- **Personal Harm:** damage to someone's reputation or mental health
- **Social Harm:** fear, conflict, or discrimination
- **Democratic Harm:** misleading voters or undermining trust in institutions
- **Safety Harm:** scams, fraud, or false medical information



Types of Harm

Even if sharing feels harmless, the impact often is not. Watch “Bad science: AI used to target kids with disinformation on YouTube - BBC World Service”



CLICK TO
VIEW

06



Exercises and
examples

(25 Min)



Exercise: Role Play – The Deep Fake Dilemma



Scenario 01

A deepfake of a classmate is shared on social media as a joke — what should friends do?

Scenario 02

A popular online influencer posts a video with their partner — later, it's revealed the partner doesn't exist and was created entirely with AI to boost views. Fans feel betrayed.

Exercise: Role Play – The Deep Fake Dilemma



Scenario 03

After a messy breakup, someone shares a video online that appears to show their ex saying cruel things or admitting to cheating. The clip spreads quickly — but it's actually an AI-generated deepfake.

Scenario 04

You discover someone used AI to imitate your voice in a prank video —
How do you handle it?

"Pope in a puffer"



Fake photos of Pope Francis in a puffer jacket go viral, highlighting the power and peril of AI

CBS News



"Pope in a puffer"



- **What happened:** An AI image of Pope Francis in a white puffer jacket went viral; many thought it was real.
- **What misleads:** Photorealism, fashion trend cues.
- **Intent:** Playful.
- **Harm:** Erodes trust, normalises not checking sources.

Which tiny clues could you check in this image before sharing?

TikTok “Bold Glamour” filter

TikTok's Bold Glamour beauty filter
worries AI experts about reality

The Washington Post



TikTok “Bold Glamour” filter



- **What happened:** A hyper-real beauty filter uses ML to reshape faces live.
- **What misleads:** Seamless AI face edits feel “natural.”
- **Intent:** Entertainment/engagement.
- **Harm:** Unrealistic standards, body-image pressure.

How could this filter affect how people see themselves and others?

“Deepfake CFO on a Zoom”

Company worker in Hong Kong pays out
£20m in deepfake video call scam

Hong Kong | The Guardian



“Deepfake CFO on a Zoom”



- **What happened:** A finance worker in Hong Kong was duped by a multi-person deepfake video call of “colleagues,” sending ~ \$25m.
- **What misleads:** Convincing video and voice clones of known staff.
- **Intent:** Fraud.
- **Harm:** Major financial loss, reputational and legal risk.

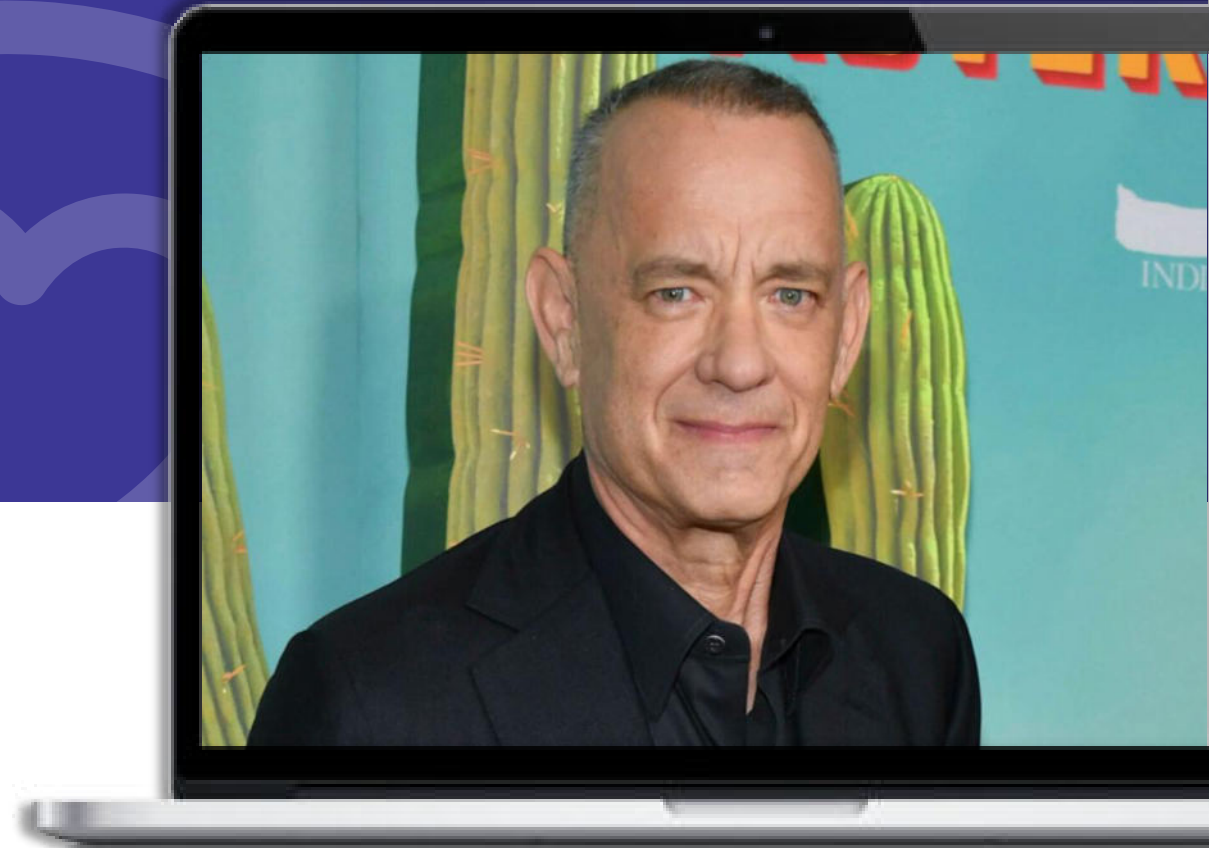
Name two out-of-band checks you would require before any transfer.

“That ad isn’t me”



Tom Hanks warns fans not to fall for
deepfake advert using his face

Ents & Arts News | Sky News



“That ad isn’t me”



- **What happened:** An AI video used Tom Hanks’ likeness to sell a product; he warned fans.
- **What misleads:** Trusted face and voice, ad format.
- **Intent:** Commercial deception.
- **Harm:** Consumer fraud, misuse of image rights.

How would you verify a celebrity endorsement before you believe it?

“Don’t vote today”



Political consultant behind AI-generated Biden robocalls faces \$6 million fine, criminal charges

AP News



“Don’t vote today”



- **What happened:** An AI voice mimicking President Biden told voters to skip the primary; fines and charges followed.
- **What misleads:** Familiar political voice, timed before voting.
- **Intent:** Voter suppression.
- **Harm:** Interferes with democratic participation.

What should a voter do on receiving a call like this?



● Exercise: Role Play – The Deep Fake Dilemma

1. Divide into 3–4 small groups.
2. Each group will act out one scenario.
3. Performance no more than
4. 2–3 minutes each.
5. Group reflection.

Case Study “The Fake Audio That Divided a Community”

CLICK TO
VIEW

The racist AI deepfake that fooled and divided a community

5 October 2024

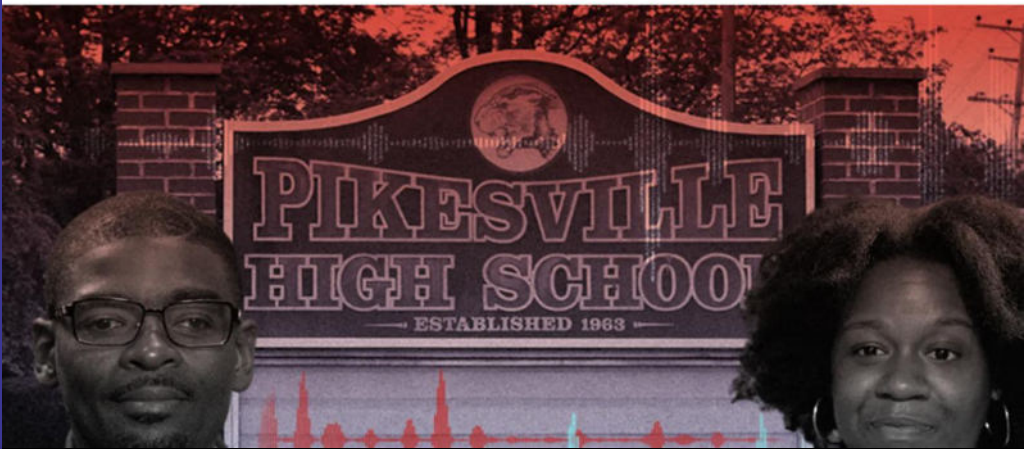
Share

Save

Add as preferred on Google

Marianna Spring

BBC Disinformation and social media correspondent



In a town near Baltimore (USA), an audio clip appeared online that sounded like a school principal making racist and antisemitic comments.

- The clip spread quickly on social media.
- Many people believed it was real.
- Parents were angry.
- The school received threats.

Case Study “The Fake Audio That Divided a Community”

CLICK TO
VIEW

The racist AI deepfake that fooled and divided a community

5 October 2024

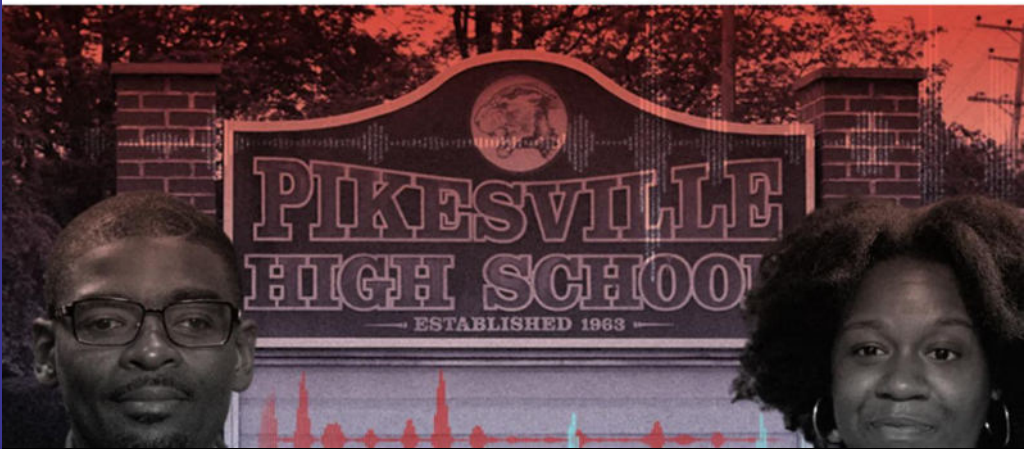
Share

Save

Add as preferred on Google

Marianna Spring

BBC Disinformation and social media correspondent



The principal was placed on leave while investigations began.

Later, police confirmed the audio was AI-generated and fake. But even after this was proven, the damage remained, trust in the school was broken, people still felt hurt and angry, some continued to believe the clip reflected “how people really think”.

Case Study “The Fake Audio That Divided a Community”

CLICK TO
VIEW

The racist AI deepfake that fooled and divided a community

5 October 2024

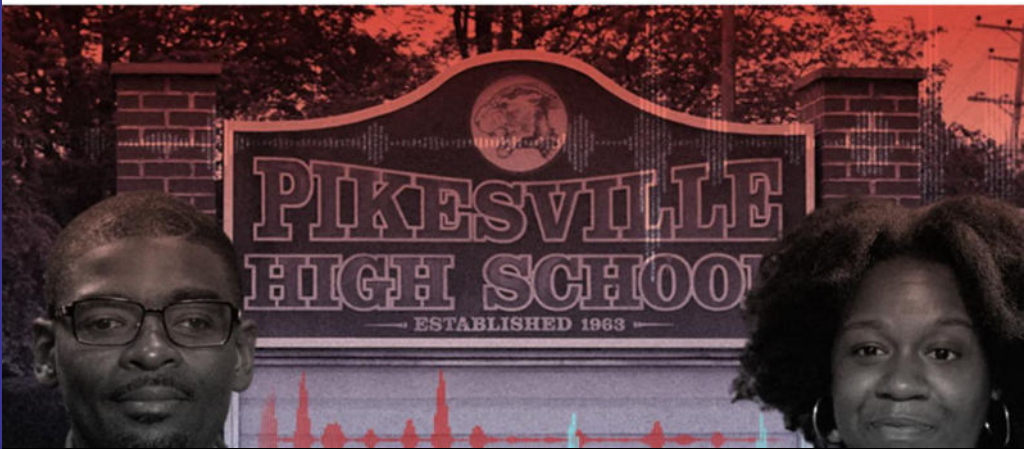
Share

Save

Add as preferred on Google

Marianna Spring

BBC Disinformation and social media correspondent



The principal was placed on leave while investigations began. Later, police confirmed the audio was AI-generated and fake. But even after this was proven, the damage remained, trust in the school was broken, people still felt hurt and angry, some continued to believe the clip reflected “how people really think”.

Case Study “The Fake Audio That Divided a Community”

Why people believed it?

Researchers explain that people believed the clip because:

- It sounded authentic (voice cloning)
- It included insider details (school language, staff names)
- It matched existing fears and experiences
- Audio-only content has fewer visual clues
- Platforms amplified it before verification

This is known as authenticity bias, when something feels real, we accept it as real (Vosoughi et al., 2018; Lewandowsky et al., 2012). AI fakes don't just trick our eyes or ears. They exploit trust, emotion, and past experience. Critical thinking is not just about spotting technical errors. It is about recognising how emotions and beliefs shape judgement.

Case Study “The Fake Audio That Divided a Community”

What Could Have Slowed the Harm?

Ask learners to discuss...

- At which moment could the spread have slowed?
- before sharing? before reacting emotionally? before platforms amplified it?
- Which THINK questions matter most here?
 - True? (Was the source verified?)
 - Intent? (Who sent it and why?)
 - Kind? (Who could be harmed?)
- Why was “it feels true” more powerful than “it is true”?



Evidence Snapshot, your “Detective Record”

- You have 5-8mins for this activity, on your own or in pairs;
- Choose one piece of content (image, video, audio, or chatbot answer).
- Create a short “Detection Note” using the template below.
- Detection Note Template (fill in)
- What is the content? (1 sentence)
- Where did I find it? (platform/account/link if possible)
- My first reaction: (what did I feel?)



Evidence Snapshot, your “Detective Record”

My checks (tick what you used)

- ☐ Details check (what looked off?)
- ☐ Source check (who posted? any trusted sources?)
- ☐ Emotion check (was I being pushed to react?)

My decision

- ☐ I trust it enough to use/share
- ☐ I’m unsure (I will not share yet)
- ☐ I think it’s misleading / fake

One sentence why (your reasoning):

Why this matters (short theory line): *(Writing your reasoning makes your thinking visible, this is a core part of critical thinking and metacognition (thinking about how you think)).*



www.valuesai.eu

Follow Our Journey



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

MODULE 04

Reliable AI-Assisted Content

Sources, Verification and Hallucination Control



Contents

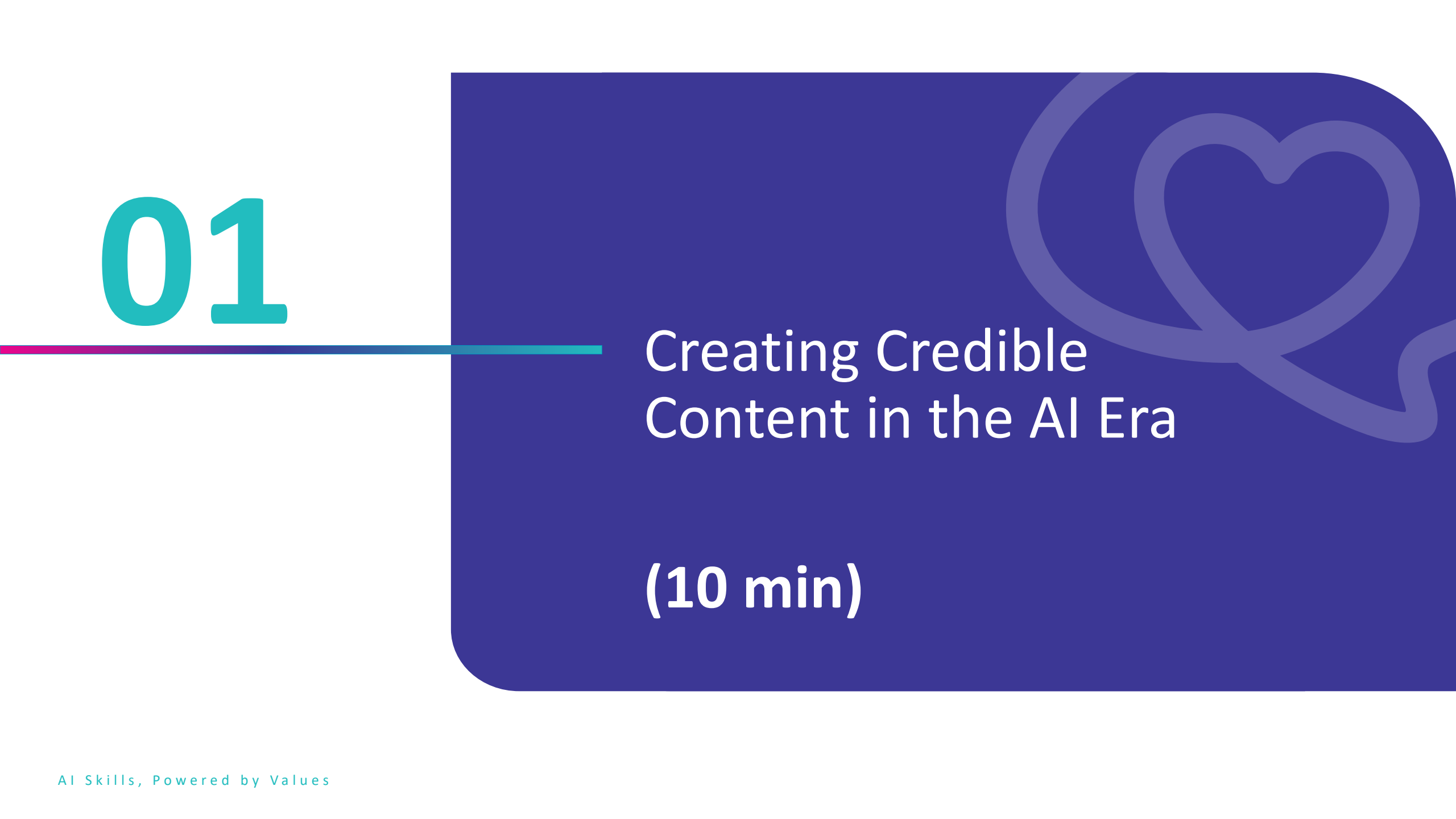
01	Creating Credible Content in the AI Era (10 min)	3
02	AI Content Risks (20 min)	8
03	Accountable Author (20 min)	15
04	Working with Sources (15 min)	23
05	AI Hallucinations (10 min)	36
06	Content Verification (10 min)	45
07	Data Analysis with AI (10 min)	51



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

01



Creating Credible Content in the AI Era

(10 min)

Reliable AI content

Creating credible content in the AI era is not only about producing text quickly, but above all about ensuring accuracy, transparency, and responsibility. Using AI does not automatically reduce quality, but it does require stronger human oversight, especially when the content concerns facts, data, health, politics, or social issues.

Reliable AI Content - Key principles



- **Cite sources.** Refer to studies, reports, official statements, and original data.
- **Verify before publishing.** Every piece of content should be checked for facts, numbers, and context.
- **Disclose AI use.** Transparency increases audience trust.
- **Write clearly and precisely.** Avoid vague wording that can lead to misinterpretation.
- **Update content.** When new data appears or an error is found, revise the material.
- **Ensure clear authorship and accountability.** Content should have an identifiable author or editorial owner.
- **Do not manipulate the audience.** AI should not be used to deliberately mislead people, even if it increases reach.

Reliable AI Content - What Builds Trust



Credible content usually explains not only what but also why and on what basis, so combining data with expert commentary and practical examples works well.

- Publication and update dates,
- Links to primary sources,
- An expert or author with relevant credentials,
- A logical text structure,
- A clear distinction between fact, opinion, and interpretation,
- No sensational or clickbait-style language.

In practice, credible content in the AI era is created when AI supports the work, but humans remain responsible for truthfulness and meaning.

Why “reliable content” matters with AI

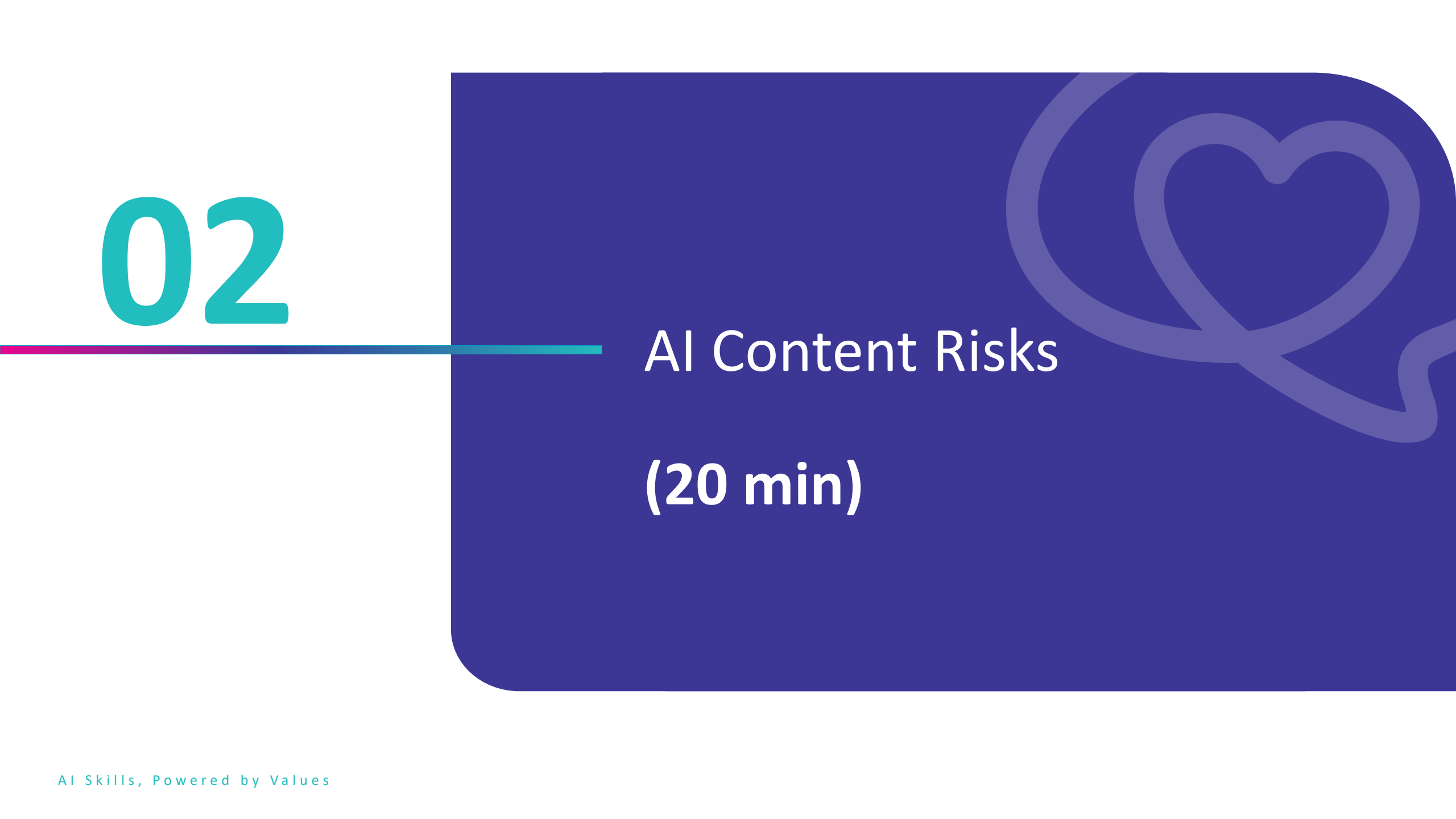
AI tools can produce content quickly, but every output is a draft: fluent text, plausible numbers, and convincing summaries that may still be wrong.

Why this matters:

- Once published, weak content can damage your credibility and mislead your audience.
- AI can fabricate citations, statistics, and quotes that look real (hallucinations).
- “Reliable” means more than fluent-it means sourced, verified and accountable.

Example: Asking ChatGPT to define a concept gives a fluent answer with no sources. The same prompt in NotebookLM, with academic PDFs loaded, returns a sourced definition tied to real authors.

02



AI Content Risks

(20 min)

AI Content Carries Three Risks:

Hallucinations,
Missing Sources,
& AI Detection

When people use AI for content, they often focus only on speed. In practice, AI output carries three major risks.

01

Hallucination Risk (False Facts):

AI can invent dates, statistics, names and quotes that sound real but never existed.

02

Source Risk (No Foundation):

General models like ChatGPT 4.0 produce fluent text without grounding it in real publications. Bibliography is missing, vague, or fabricated.

03

Detection Risk (AI Fingerprints):

Detectors flag generic, list-like, formulaic AI text. Output without editing scores 100% AI; grounded, edited output scores under 10%.

“Looks reliable” but isn’t (Myth vs Reality)



Watch for:

- Perfectly formatted citations with wrong page numbers
- Real author names attached to invented papers
- Plausible statistics with no verifiable source
- Fluent definitions that contradict the actual literature

Myth

“It’s reliable if it sounds academic.”

Fact

AI imitates academic style perfectly while still inventing facts.

Myth

“It’s safe if a detector says it’s human-written.”

Fact

Detectors only check style. Edited AI output can score 92% original and still contain hallucinations.

What is AI SLOP?

AI slop = low-quality, misleading, or inaccurate content generated by AI.



It sounds confident, but includes:

- Fabricated facts
- Wrong numbers or dates
- Incorrect names, citations, or scientific claims
- Repetitive, generic structure
- “Smooth” text with zero real evidence

Why it happens?

01

The model predicts patterns, not truth – it generates the most likely response based on data, but it does not actually “know” whether it is true.

02

It may fill in gaps with invented details – when information is missing, the model can add something that sounds logical but may not reflect reality.

03

It does not always have up-to-date information – if the model’s knowledge is time-limited, it may provide outdated facts or miss recent events.

04

It may misunderstand context – an unclear question, shorthand, or ambiguity can lead to the wrong answer.

Why it happens?

05

It sounds confident even when wrong – the wording may appear highly reliable even if the content contains errors.

06

It can mix real facts with hallucinations – part of the answer may be correct, while other parts are invented or unverified

07

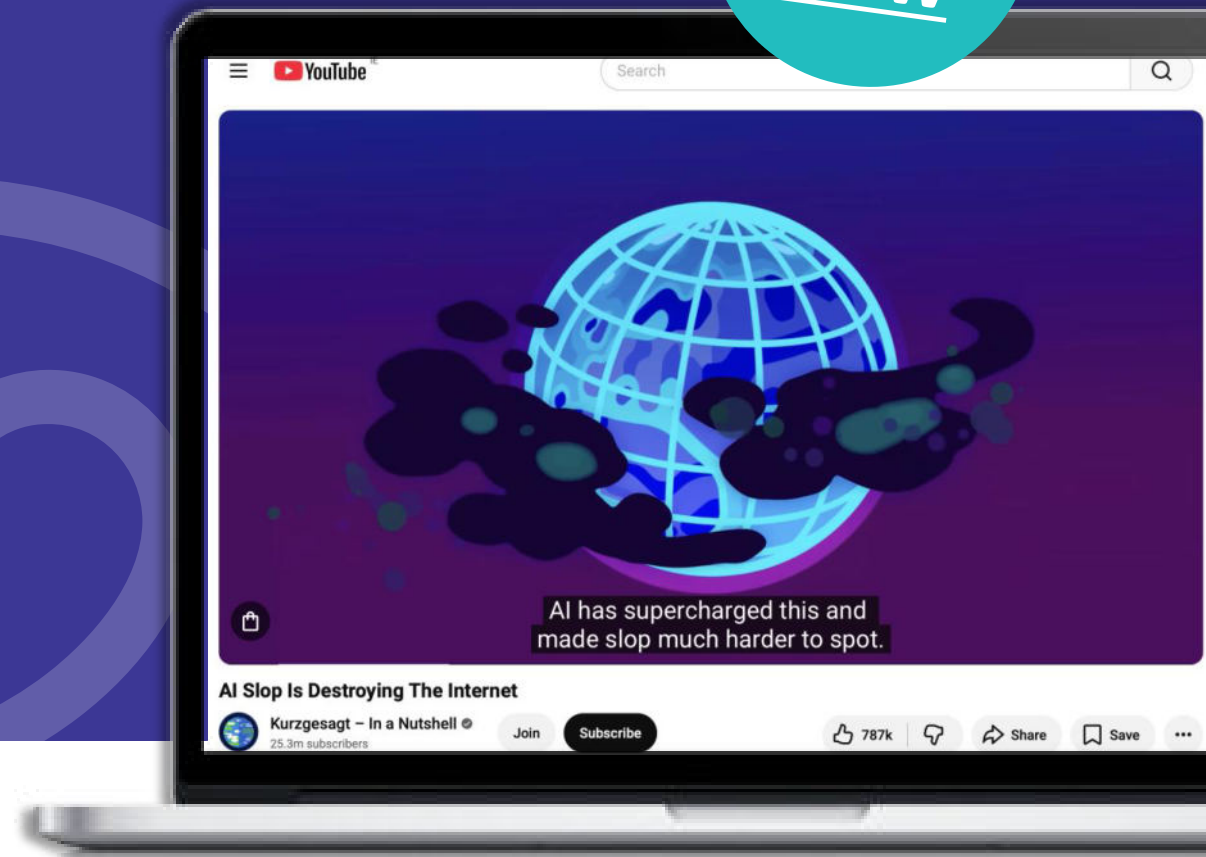
It depends on the quality of the input – if the prompt is unclear, contradictory, or incomplete, the output may be less accurate.

This is why verification becomes a key digital skill

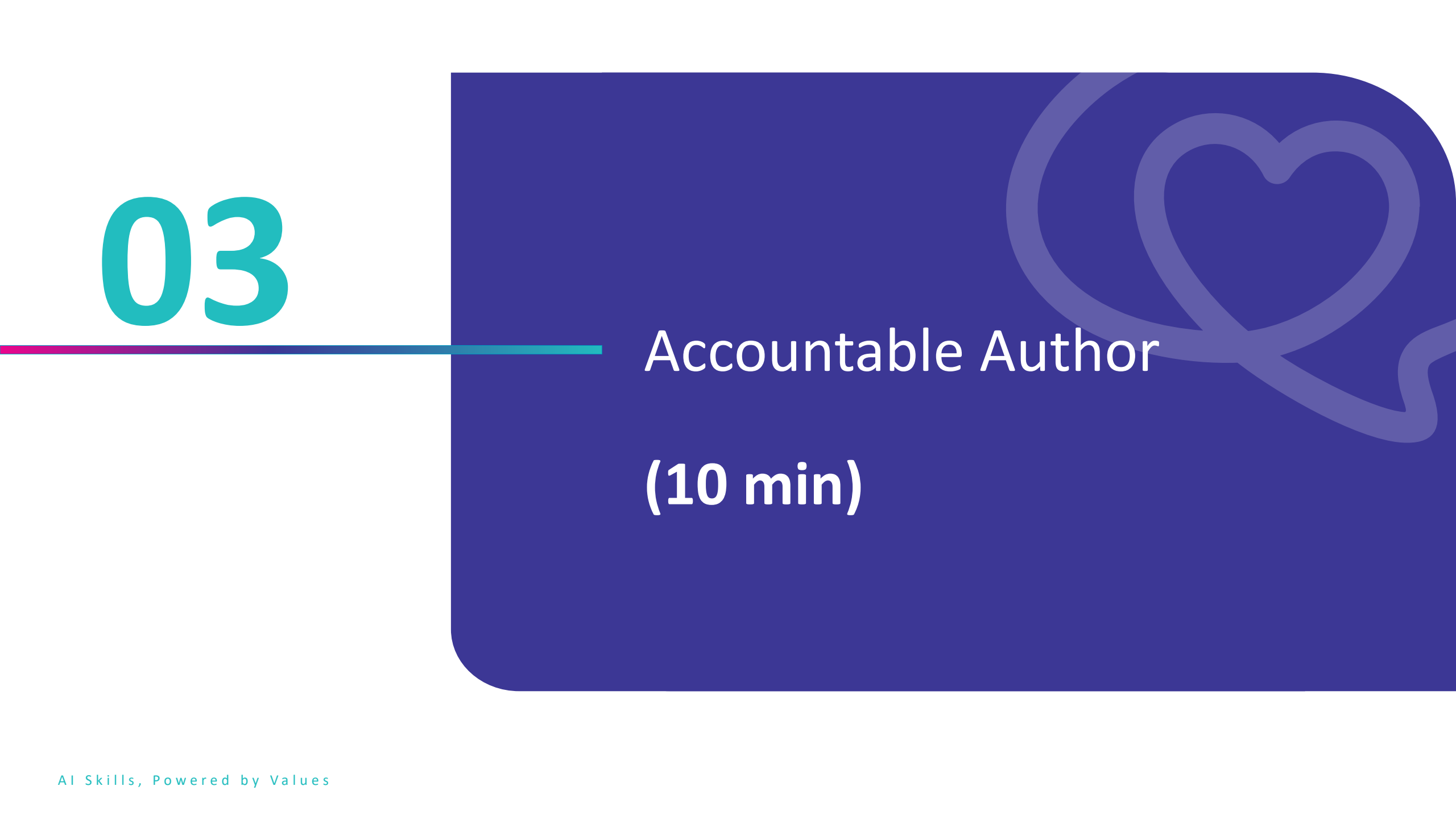
How AI SLOP destroying the Internet

AI-generated content is flooding the internet, from videos and music to news and books. On top of that, generative AI is making false information look more convincing than ever. We're entering a new era of information overload, and figuring out what's real and what's not is becoming more and more of a challenge.

CLICK TO
VIEW



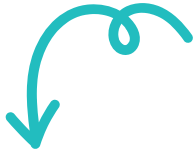
03

The background features a large, stylized heart shape in a light blue color, positioned on the right side. A horizontal line with a gradient from pink to blue runs across the middle of the slide, passing behind the text.

Accountable Author

(10 min)

Accountable Lens



Use four accountability questions whenever you produce content with AI:

Accuracy

Did I check facts, dates, and quotes against trusted sources?

Sources

Can I cite real publications behind every key claim?

Originality

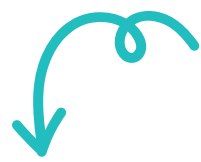
Did I edit the AI draft so it is not detectable as raw AI output?

Accountability

Can I explain how I used AI, what I changed, and why?

Reliable content = smart use of AI guided by accountability.

Accountable Lens



**Safe starts
before AI**

What you
enter or upload
matters first

**Rights and
consent matter**

Other people's
data needs
permission
and care

**Fairness is part
of safety**

Biased / low-quality
data can cause
harm

**Safer workflows
exist**

You can often
get AI help while
sharing far less

Before you use AI: every output is a draft

**Many people focus on AI speed.
Reliable content starts earlier:**
what you verify, ground, and edit.



In practice, AI tools generate:

- Definitions and explanations (often without sources)
- Summaries of papers (sometimes invented or distorted)
- Lists of references (frequently fabricated)
- Data analyses and interpretations (without methodology)

Before you use AI: every output is a draft

Reliable AI use begins with a simple habit: treat every output as a draft to verify, not a final answer.

AI output is a draft, not truth

What counts as content you must verify:

- Definitions and theoretical claims
- Statistics, percentages, and dates
- Names of authors, books, and articles
- Data analyses and conclusions



DO



DO Remember:

Every fact, citation, and number from AI must be checked before use.



DON'T



DON'T Assume:

"It sounds professional" or "It must be true if AI said so."

The Never Trust List (NTL)

Do not publish without checking:



If it cannot be traced to a verified source-do not publish

- **Citations:** any reference to a book, paper, or author the AI named
- **Quotes:** any “direct quote” the AI attributes to a person
- **Statistics:** percentages, sample sizes, study results, rankings
- **Historical facts:** dates, events, locations, names of laws or treaties
- **Definitions and theories:** claims about what a concept “is” or who proposed it
- **Numerical analysis data:** client lists, HR data, internal documents, non-public figures

How to make AI output more reliable (before you publish)



Before publishing AI output, apply these checks:

- Verify citations: open in Google Scholar, Scopus, or publisher site
- Cross-check every statistic with at least one independent source
- Re-run the prompt in a grounded tool (e.g., NotebookLM with real PDFs loaded)
- Edit structure, replace generic phrases, add specific examples

If you cannot find the source of a claim → do not publish it

04

A decorative graphic consisting of a horizontal line with a color gradient from pink to blue. To the right of the line is a large, dark blue rounded rectangle. Inside this rectangle, on the right side, is a faint, light blue heart shape. The text '04' is positioned to the left of the line.

Working with Sources

(15 min)

○ Sources Provided by AI

01

Using credible sources – choose materials from reputable journals, public institutions, scholarly databases, and expert organizations.

02

Distinguishing scholarly, industry, media, and opinion sources – each source type serves a different function and offers a different level of reliability.

03

Checking publication dates – currency matters, especially in fast-changing fields where AI may provide outdated information.

04

Verifying citations – check whether the cited author, title, year, and quoted claim actually exist and match what the text or AI says.

○ Sources Provided by AI

05

Distinguishing primary and secondary sources – a primary source presents original material, while a secondary source discusses or interprets earlier work.

06

Cross-checking information – important claims should be confirmed in at least two or three independent sources.

07

Evaluating the author, publisher, and purpose – always ask who wrote the text, where it was published, and whether its goal is to inform, sell, persuade, or comment.

Example of AI providing an incorrect source

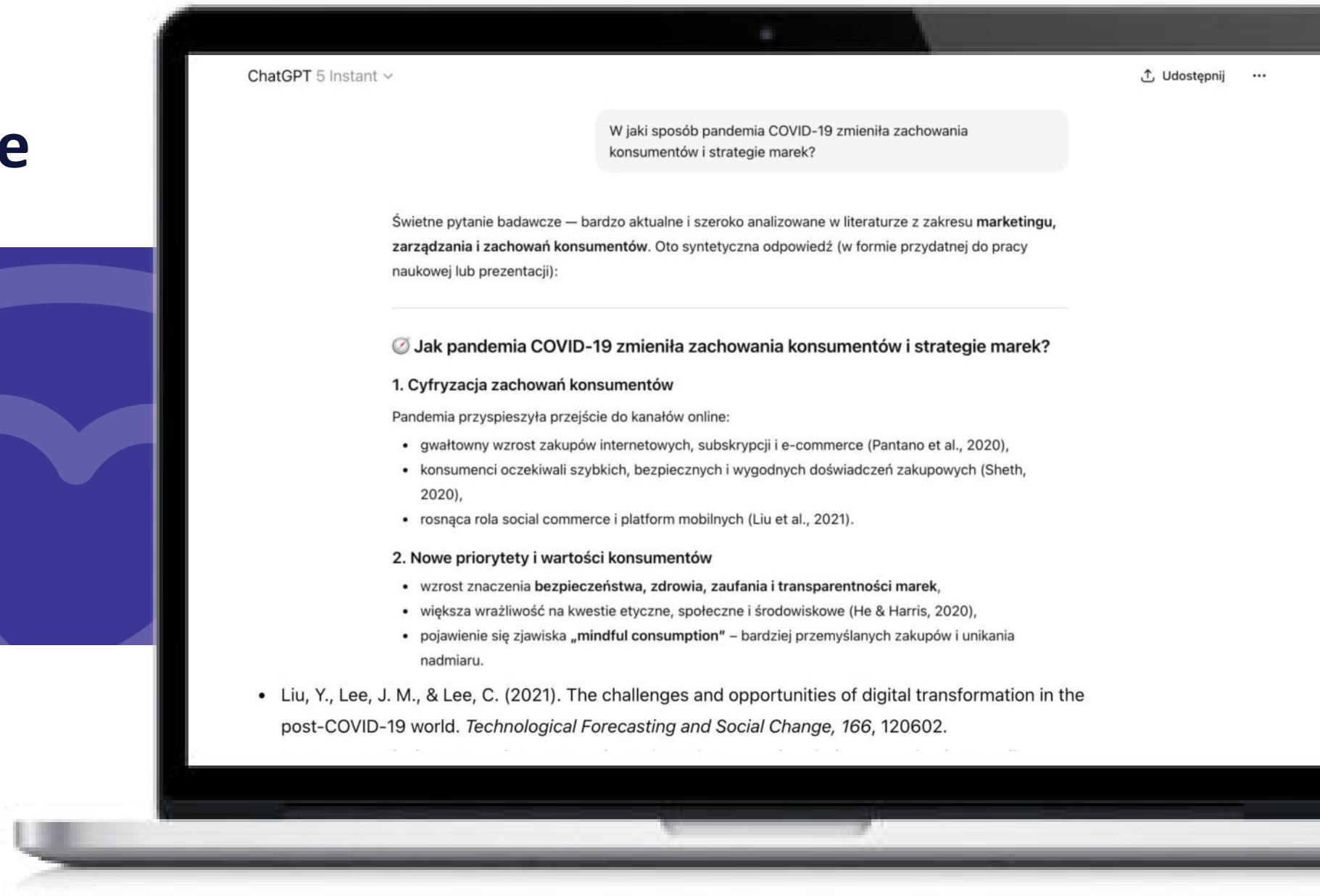
This slide shows a ChatGPT 5 Instant conversation where the user asks about how the COVID-19 pandemic changed consumer behavior and brand strategies. The AI provides a structured response with citations to academic sources (Pantano et al., 2020; Sheth, 2020; **Liu et al., 2021**; He & Harris, 2020), discussing digitalization of consumer behavior and new consumer priorities including safety, health, trust, and "mindful consumption."

When the user verified one citation, they discovered that the article „The challenges and opportunities of digital transformation in the post-COVID-19 world” **was NOT published** in Technological Forecasting and Social Change as the AI claimed, but rather in IEEE Engineering Management Review, Vol. 48(3), pp. 97-103.

See the
next slide



Example of AI providing an incorrect source

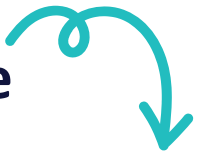


Example of AI providing an incorrect source.

Conclusion

- The AI misattributed a real article to the wrong journal, incorrectly citing it as published in Technological Forecasting and Social Change instead of IEEE Engineering Management Review.
- It generated a plausible but false bibliographic reference, including journal name and publication details, creating a misleading APA-style citation.
- The AI failed to verify publication metadata (journal, publisher, DOI) despite the source being easily verifiable.
- It combined a real title with an incorrect publishing context, a typical case of source hallucination.
- This error undermines academic reliability, potentially leading to incorrect bibliographies and issues during peer review.

See the
next slide

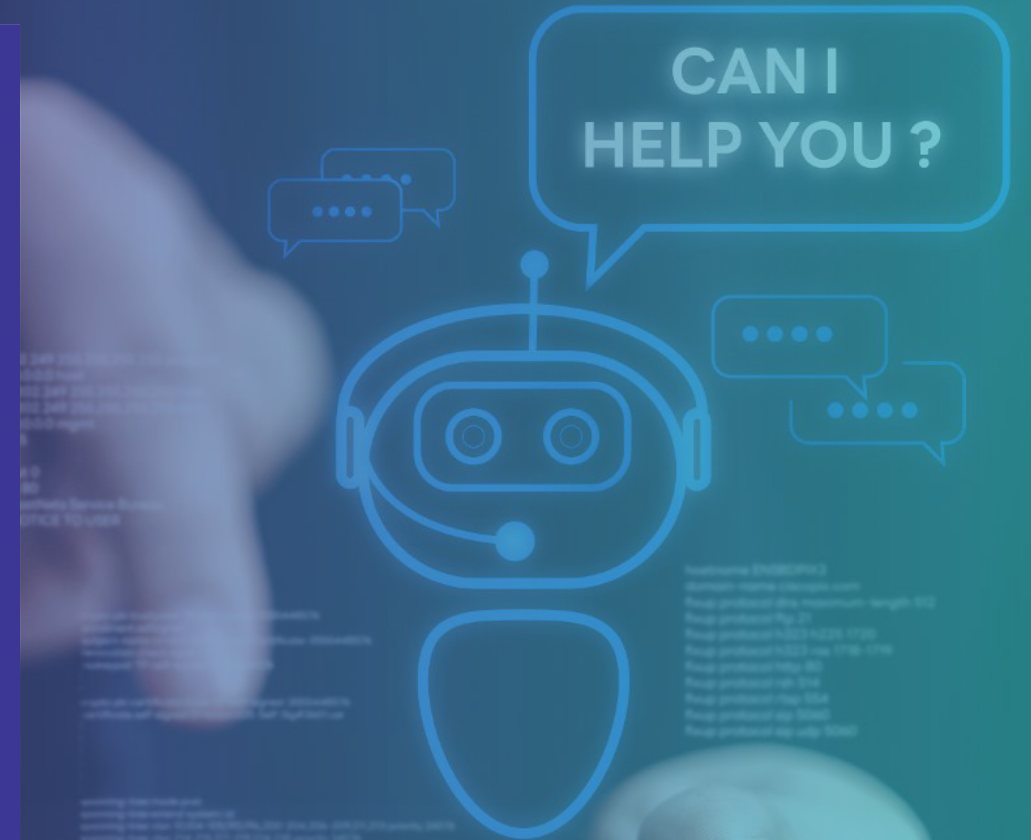


Example of AI providing an incorrect source. Conclusion

- Bhatti, A. (2020). Online shopping behavior during COVID-19 pandemic. *International Journal of Business and Social Science*, 11(10), 1-10.
- He, H., & Harris, L. (2020). The impact of COVID-19 pandemic on corporate social responsibility and marketing philosophy. *Journal of Business Research*, 116, 176-182.
- Huang, M.-H., Rust, R. T., & Maksimovic, V. (2021). The feeling economy: Managing in the next generation of artificial intelligence. *California Management Review*, 63(3), 43-65.
- Liu, Y., Lee, J. M., & Lee, C. (2021). The challenges and opportunities of digital transformation in the post-COVID-19 world. *Technological Forecasting and Social Change*, 166, 120602.
- Pantano, E., Pizzi, G., Scarpi, D., & Dennis, C. (2020). Competing during a pandemic? Retailers' ups and downs during the COVID-19 outbreak. *Journal of Business Research*, 116, 209-213.
- Sheth, J. (2020). Impact of COVID-19 on consumer behavior: Will the old habits return or die? *Journal of Business Research*, 117, 280-283.

Example of AI providing an incorrect source. Conclusion

AI can effectively summarize literature, but all AI-generated references **must be independently verified** before use in academic work.





● Exercise

Ask any AI model to generate a one-page A4 write-up on any topic. Ask it not to use websites, but instead to rely on publications such as articles, books, reports, and similar sources

- Create a list of sources.
- Then, according to good practice, verify the accuracy of the sources provided: do they really exist?
- Are they accessible?
- Do they list the correct authors? .

If you detect an AI mistake, share your observations with the group and the instructor.

Scenario: ChatGPT 4.0 vs NotebookLM

A research case study compared two tools on the same task: *“Generate a definition of innovation strategy with a bibliography.”*

The same prompt produced two outputs that show what grounding changes:

- ChatGPT 4.0: fluent definition, no real bibliography, detected as 100% AI by Originality.ai.
- NotebookLM (with real PDFs): grounded definition citing real authors (Porter, Janasz, Pomykalski, Freeman); 92% original score.

What this means in practice:

- Without grounding, AI invents fluent but unsourced content. With a real knowledge base, the same model produces verifiable text tied to specific publications.

ChatGPT 4.0 vs NotebookLM: what grounding changes

Sources are not just academic - they are what make AI output accountable.

ChatGPT 4.0 (no grounding):

- Fluent text, simplified definitions, no bibliography
- Plausible company examples (Google, Tesla, Apple) with no „studies” behind them.
- 100% detected as AI-generated by Originality.ai

I m NotebookLM (with PDFs loaded):

- Theory-grounded definitions citing Porter, Janasz, Pomykalski, Freeman
- Real research traditions: innovation typologies, network economy, SME context.
- Concrete bibliography in APA format, ready to verify
- Style scored 92% original by Originality.ai

Sources and Rights:

“Do I have a real reference for this claim?”

Reliable AI use is not only about facts. It is also about authorship: who wrote the original idea, whose work you cite, and how you credit it.

You can usually rely on AI cannot trust AI output for:

- Exact statistics, percentages, sample sizes
- Specific quotes attributed to named authors
- Publication years, page numbers, journal titles
- Definitions of niche or recent concepts not in mainstream literature
- General explanations of well-known concepts
- Structural editing (grammar, flow, paragraphing)
- Summarising text you have provided to AI yourself

What a “real reference” looks like (and what it does not)

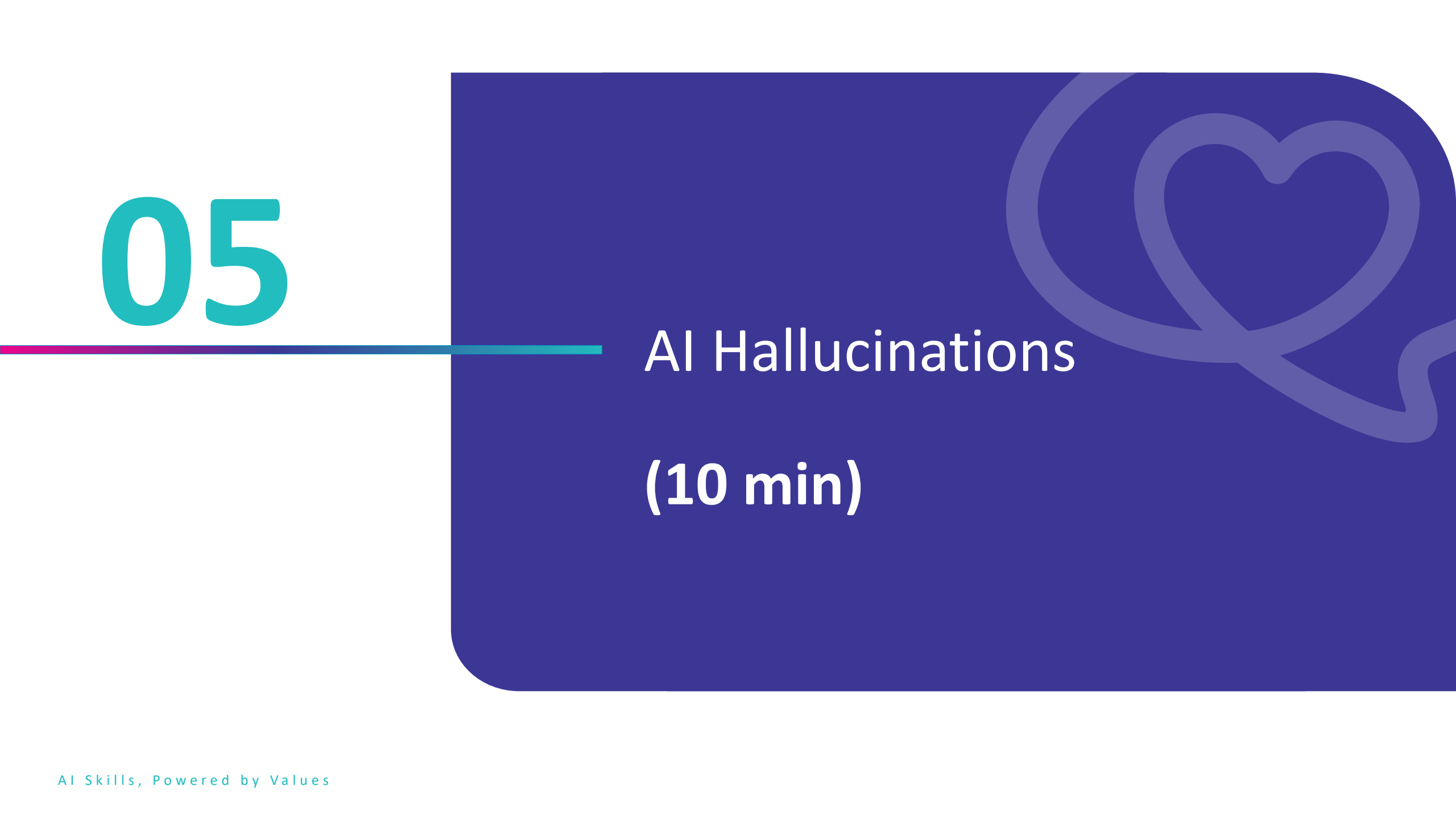
Reliable AI use is not only about facts. It is also about authorship: who wrote the original idea, whose work you cite, and how you credit it.

A real source is not only a name and a year. For AI-cited references, a source must be:

- **Real:** the publication actually exists and is findable in a database
- **Specific:** matches the claim word-for-word, not just the topic
- **Accessible:** you can open it and read the relevant section
- **Authoritative:** published by a credible journal, publisher, or institution
- **Important:** Consent depends on audience and tool

A person might agree to tell you something, but not agree to have it shared with an AI tool or stored in a system.

05



AI Hallucinations

(10 min)

What AI hallucinations are?



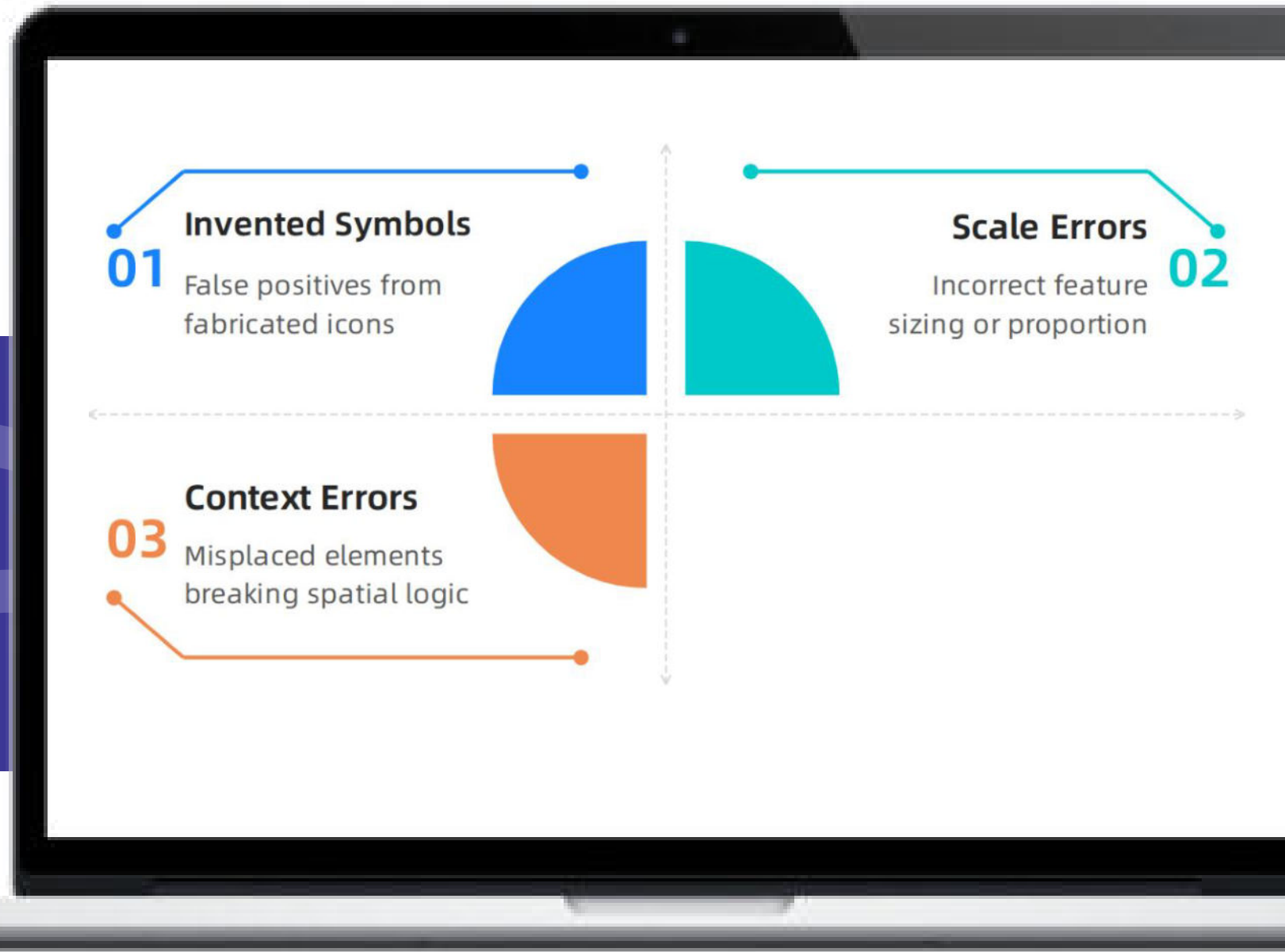
AI hallucinations are model-generated answers that sound plausible but contain false, invented, or misleading information. The model does not „know” what is true - it usually predicts the most likely text rather than checking reality.

AI hallucinations occur when a system produces content that appears correct but is factually wrong, such as a **fabricated quote, a nonexistent source, an incorrect date, or a misleading explanation**. These are not medical hallucinations — they are information-generation errors, not perceptual experiences.

AI Hallucinations Map

How to detect and fix AI hallucinations in graduated symbol

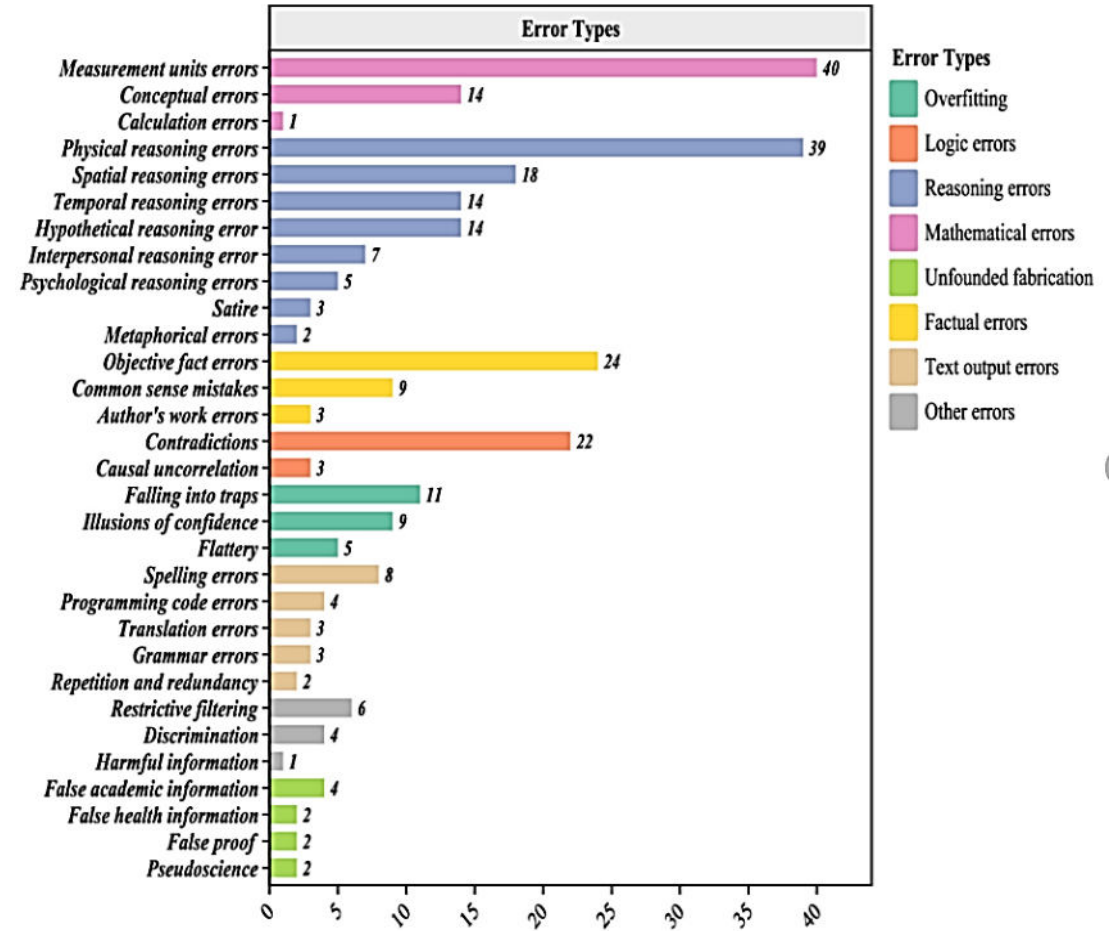
maps. Aihallucinationreport.com. <https://aihallucinationreport.com/how-to-detect-and-fix-ai-hallucinations-in-graduated-symbol-maps/>



AI Hallucinations

Statistics for 284 analysis units.
Different colors represent
different error types.

Sun, Yujie & Sheng, Dongfang & Zhou, Zihan & Wu, Yifei.
(2024). AI hallucination: towards a comprehensive
classification of distorted information in artificial intelligence-
generated content. *Humanities and Social Sciences
Communications*. 11. 10.1057/s41599-024-03811-x.



○ Types of AI Hallucinations

01

Factual Hallucinations – the model states false facts, events, people, or definitions.

02

Source Hallucinations – AI invents sources, publications, authors, titles, or assigns claims to nonexistent texts.

03

Quote Hallucinations – the model gives a quote that was never actually said or falsely attributes it to a person.

04

Numerical Hallucinations – AI miscalculates data, confuses percentages, dates, statistics, or table values.

Types of AI Hallucinations

05

Contextual Hallucinations – the model misunderstands the question or adds details the user did not intend.

06

Mixed Hallucinations – the answer contains both correct facts and invented elements, making the error harder to spot.

Causes of AI Hallucinations

01

The model predicts patterns, not truth – it generates the most likely words rather than checking facts.

02

Training data may be incomplete or flawed – if the data contain gaps, errors, or contradictions, the model can reproduce them.

03

Lack of up-to-date information – the model may not know recent events or changes.

04

Unclear prompt context – with ambiguous or shorthand prompts, the model may misread the user's intent.

Causes of AI Hallucinations

05

Systems reward fluent answers – models are often trained to respond confidently and smoothly, even when uncertain.

06

No built-in fact checking – without access to external sources, the model cannot verify whether its answer is true

Indirect Hallucinations: when names and dates look right but aren't

Common Indirect Hallucinations:

- Real author + real journal + invented paper title
- Real paper + wrong year, page, or volume number
- Real concept + wrong author (“porter said” when schumpeter did)
- Real statistic + invented context (right number, wrong study)

Safer Alternatives:

- **Search:** open google scholar and confirm the exact paper exists
- **Replace:** cite a paper you have actually read on the same topic
- **Soften:** use general phrasing if you cannot pin down a specific source
- **Remove:** if you cannot find the source, delete the claim entirely

AI can hallucinate even when names, dates and titles look correct.
The pattern is plausible but the link to the original work is broken.

06

Content Verification (10 min)

AI-generated content can be useful, but it should not be published or used without prior verification



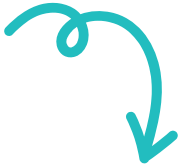
Treat the AI answer as a draft, not as final truth.

Check claims, numbers, quotes, and sources before using them.

Document what was verified and on what basis.

Step 1:

Identify Risk & Key Claims



First, identify which parts of the AI response require **mandatory checking**: facts, dates, names, numbers, quotes, definitions, and cited sources. The greater the impact on health, law, education, finance, or reputation, the stricter the verification should be.

Which elements could cause harm if they are wrong?

Did the AI provide real sources, or does it only sound convincing?

What in this answer is fact, and what is interpretation?

Step 2: Verify Facts & Sources

Every important claim should be checked against credible external sources, preferably **primary ones**: scholarly articles, reports, public institution websites, and reputable industry publications. Never assume that a source cited by AI actually exists — verify the author, title, year, accessibility, and whether the source truly contains the stated information.

Confirm the information in at least two independent sources.

Open the original source, not only a summary or repost.

Check the publication date and currency of the information.

Step 3: Assess Quality, Language, & Context



After checking the facts, assess whether the text is logical, clear, balanced, and appropriate for its purpose. AI may produce content that is formally correct but poorly contextualized, overly confident, biased, or stylistically unsuitable.

The text mixes facts with opinion.

It contains bias, oversimplification, or excessive generalization.

The tone and style fit the audience and publication goal.

Edit, Decide, & Document

Finally, correct errors, remove uncertain fragments, and decide whether the content is fit for use. Good practice is to record what was changed, what was rejected, and which sources confirmed the final version of the text. **Final decision:**



Publish



If the content has been checked and corrected.



Hold



If some information remains uncertain.

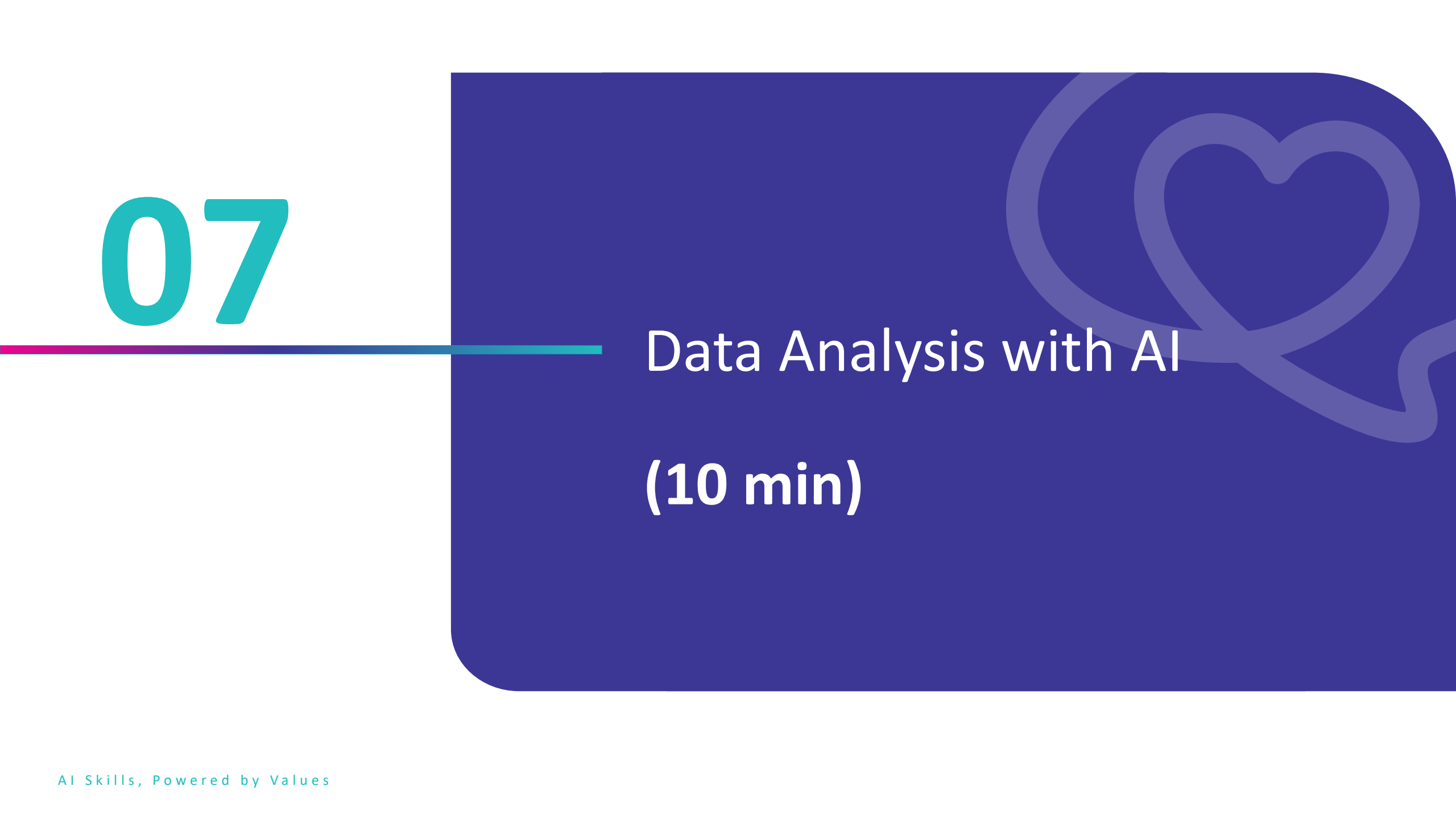


Reject



If the content contains too many errors or lacks source confirmation.

07



Data Analysis with AI (10 min)

When you analyse data with AI, validation is the work

When you ask AI to analyse data, the system will generate a fluent narrative. Depending on prompt and tool, the output may also:



- Skip the methodology (no mention of sample size, method, or assumptions)
- Average over outliers without flagging them
- Generalise from a small sample (e.g., n=10 respondents)
- Produce confident conclusions (where data only supports tentative ones)
- Confirmation bias: AI repeats the framing of your prompt back to you

When you analyse data with AI, validation is the work

Why validating analysis matters:



- You catch errors before they reach a report or publication.
- You distinguish strong findings from confident-sounding noise.
- You avoid drawing decisions from unrepresentative samples.
- You keep authorship of the analysis - not the model

SAFE Analysis

Ladder: choose the most verifiable level

When you analyse data with AI, do not start at the top of the ladder. Choose the lowest level that still answers your question.

**Stop rule: If you cannot remove sensitive/confidential elements
→ do not upload**

LEVEL 1

(Most reliable): Ask a general question (no personal data).

Example: "Explain when a t-test is appropriate vs Mann-Whitney U."

LEVEL 2

Use an anonymised summary (roles, no identifiers).

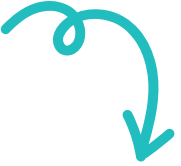
Example: "Mean = 4.2, SD = 1.1, n = 120 - what does this suggest about the trend?"

LEVEL 3

Use limited, de-identified data only if necessary.

Example: "Here is the dataset. Compute X, then I will recheck against my own calculation."

Conclusion from Small Data



Unreliable AI output (realistic):

“AI ran sentiment analysis on 200 tweets about our brand. 67% positive. Therefore brand sentiment is strongly positive.”

What is unreliable?

- “67%” with no methodology, no model, no validation
 - 200 tweets is a tiny, possibly biased slice
- “strongly positive” overstates what the number shows



www.valuesai.eu

Follow Our Journey



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

MODULE 05

Data Privacy & Safe Use of AI Tools

Share less. Protect more.
Use AI responsibly.

Contents

01	Why Safe Data matters with AI (15 min)	3
02	How to share data safely? (15 min)	10
03	Privacy, rights, and consent (15 min)	16
04	Choosing safer AI workflows (25 min)	23
05	Some Data (15 min)	33



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

01



Why Safe Data
matters with AI

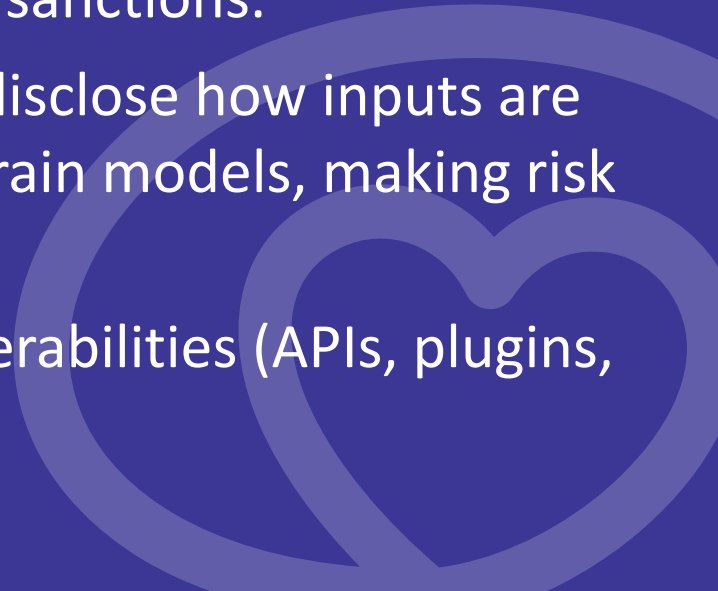
(15 min)

Data and AI

Sharing data with AI at work is risky because it can cause data exposure, regulatory breaches, and reputational or operational harm if inputs are stored, reused, or leaked by third-party models.

Key Risks



- **Sensitive-data exposure:** prompts or files sent to AI platforms can be retained or appear in model outputs, leaking intellectual property, customer PII, or financial records.
 - **Shadow AI and uncontrolled use:** employees using unapproved AI tools create blind spots where data leaves corporate control and vendors may keep or reuse inputs.
 - **Legal and compliance breaches:** using external AI services without proper contracts or data-processing agreements can violate privacy laws and lead to sanctions.
 - **Lack of transparency and auditability:** many AI providers don't disclose how inputs are stored, where they are processed, or whether they are used to train models, making risk assessment difficult.
 - **Expanded attack surface:** integrating AI increases technical vulnerabilities (APIs, plugins, endpoints) that attackers can exploit to access data.
- 

Three negative effects: privacy, unfairness, and false confidence

01

Privacy Risk (Exposure):

Oversharing can reveal personal, sensitive, or third-party information-sometimes even indirectly.

02

Fairness Risk (Harm):

If the data is biased, incomplete, or uses unfair labels, AI outputs can reinforce stereotypes or disadvantage groups.

03

Reliability Risk (False Confidence):

AI can produce fluent answers that sound correct even when the input data is weak, missing context, or misleading.



DO



Ask: What is missing?
Who is not represented?
Could this output harm
someone?



Use checks: Consent
Gate + Fair Data Check +
SAFE Workflow Ladder



DON'T



**Treat AI outputs as
neutral truth**



**Use AI results to judge
people or make decisions
without context**

Values Lens

Safe data = smart choices
guided by values.

Use four value questions whenever you use
AI with data:

1. **Agency:** Am I choosing deliberately, or clicking automatically?
2. **Privacy:** Am I sharing more data than necessary (about me or others)?
3. **Fairness:** Could my use of data harm or disadvantage someone?
4. **Transparency:** Can I explain what data I used and why?

Special categories of personal data

Particular caution should be exercised when sharing them with AI systems.

1. Personal Data
2. Sensitive Data
3. Health Data
4. Financial Data
5. Login Credentials
6. Customer Data
7. Internal Organizational Documents
8. Unpublished Research Results
9. Data Protected By Trade Secrets

02

How to share
data safely?

(15 min)

Before you use AI: your input becomes data

Many people focus on what AI answers.

Safe data starts earlier:
what you type, paste, or upload.



In practice, AI tools can work with:

- Your text (prompts, messages, copied content)
- Your files (documents, spreadsheets, pdfs)
- Your images (photos, screenshots)
- Your context (details that allow inference)

Before you use AI: your input becomes data

If you share too much, you may create risks you cannot fully undo: privacy exposure, harm to others, and unfair or misleading outputs.

Safe AI use begins with a simple habit: share the minimum necessary; or do not share at all.



DO



DO Remember:

Anything you type, paste, or upload is data you share.



DON'T



DON'T Assume:

"It's only text" or "It's just a quick upload."

“Looks harmless” but isn’t (Myth vs Reality)



Myth

“It’s safe if I remove the name.”

Fact

People can be identified indirectly through combinations like:

- age + city + school/internship
- unique event details
- screenshots with small visible clues
- location/time patterns

Myth

“It’s safe if it’s only a screenshot.”

Fact

Screenshots often include names, faces, contact info, or context you miss.



● Micro-task: Spot the red flags

Task (60–90 seconds):

Mark each item as Never / Avoid / Usually OK:

- A password reset code
- A CV with full address and phone number
- A chat screenshot with names visible
- A general question: „How do I structure a CV?”
- A school list with names and emails
- A survey summary with totals only (no names)

You Never Know

- It is not always clear where data are processed
- Data may be stored in the conversation history
- They may be used to improve services, depending on the settings and terms of service
- The user may accidentally disclose confidential information



03



Privacy, rights,
and consent

(15 min)

Decide before you share data about others

Use this gate whenever your input includes other people's information (messages, photos, lists, stories, documents).

Ask 3 questions:

- **Right/permission:** Do I have permission (or a clear right) to use this data with AI?
- **Risk:** Could this identify, embarrass, or harm someone now or later?
- **Safer alternative:** Can I achieve the goal with less data (or no personal data)?

Rule

If you answer No / Not sure to any question → do not upload the data. Choose a safer workflow (e.g., ask a general question without private details). **“Not sure” counts as a warning sign**

Consent and other people's data: what responsible use looks like.

Privacy and rights are not only technical issues-this is about respect and preventing harm.



DO (safer alternatives):

- Ask for consent and respect "no."
- Replace names with roles: Person A / friend / teacher / colleague.
- Remove unique details (exact places, dates, identifiers).
- Ask a general question: "How can I support a friend who is stressed?" (no personal story)

DO



DON'T (high risk):

- Paste someone else's private messages into AI „for advice.”
- Upload group lists with names/emails to „organise” them.
- Share personal stories about others without asking

DON'T



Quick practice: Does it pass the Consent Gate?

Instructions

For each scenario, decide: Pass (safe enough) or Do not pass (use a safer option). Then write one short reason.

- “Summarise my teacher’s feedback letter (it includes the teacher’s name and school).”
- “Help me write an apology message. I will describe the situation without pasting the full chat.”
- “Analyse a volunteer list with names and phone numbers to find ‘unreliable’ people.”

Do I have the right to use this data?

Safe AI use is not only about your privacy. It is also about rights: whether you are allowed to use, share, or process data (especially when it belongs to someone else or to an organization).

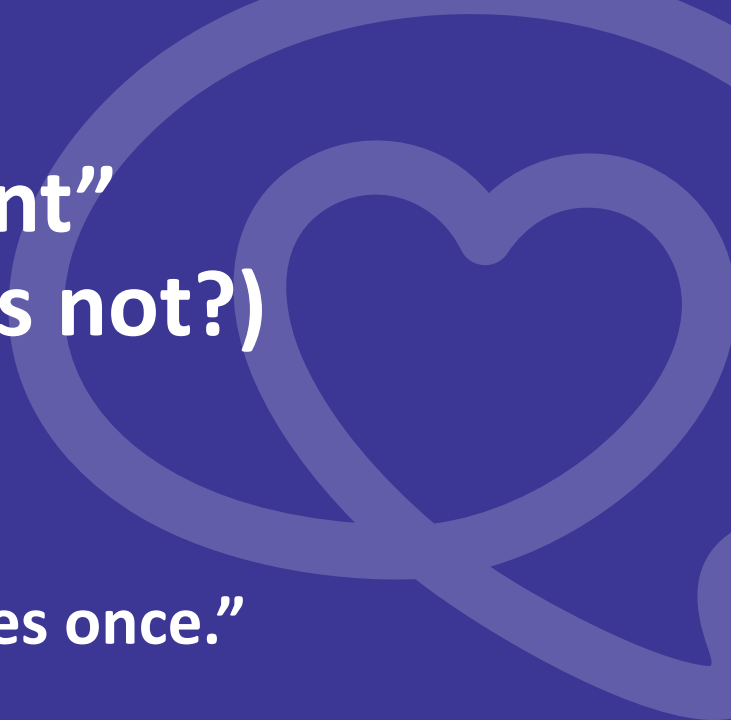
You usually do not have the right to share with AI:

- Private messages of other people (even if you are part of the chat)
- Lists of people (class lists, volunteer lists, client lists)
- Documents that are not yours (contracts, HR notes, internal emails)
- Personal stories told to you in confidence (health, relationships, family issues)

You usually do have the right to use:

- Your own text that contains no sensitive details
- General questions without personal data
- Anonymised examples you created yourself (fictional or de-identified)

Consent: what “real consent” looks like (and what it does not?)



Consent is not only “someone said yes once.”
For AI use, consent should be:

- **Informed:** the person understands what will be shared and why
- **Specific:** consent for this purpose, not “anything you want”
- **Voluntary:** no pressure, no fear of consequences
- **Revocable:** the person can change their mind
- **Important:** Consent depends on audience and tool
- A person might agree to tell you something, but not agree to have it shared with an AI tool or stored in a system.

Indirect Identification: When “removing names” is not enough?

People can be identified even without names if details allow someone to guess who it is

Common indirect identifiers:

- Exact age, exact location, school/ employer, unique role exact dates and events (“the only person who...”)
- Combinations (age + city + hobby + situation)
- Screenshots with small visible clues (profile pictures, group names)

Safer alternatives (choose one):

- **Abstract:** describe the situation without unique details
- **Generalise:** use ranges (age group), broad areas (region)
- **Role-based:** Person A / Person B instead of real identities
- **Do not upload:** if you cannot remove identifying context

04



Choosing safer AI workflows

(25 min)

Want to Choosing safer AI workflows - same goal, less data

Many risks with AI do not come from the tool itself, but from how we use it:we often share too much data because it feels faster or easier.

Want to Choosing safer AI workflows - same goal, less data

When you share data with AI, it can be processed, stored, and repeated



When you type, paste, or upload data to an AI tool, the system must process it to generate an answer. Depending on the service and settings, your data may also be:

- Stored or logged (for service improvement, troubleshooting, security, or compliance)
- Reviewed in some cases (e.g., to improve safety or quality)
- Used to personalise your experience (in some tools)
- Reflected back in outputs
- Shared indirectly when you copy/paste outputs into other apps or send screenshots

Want to Why sharing less data matters?

- You reduce the chance of privacy leaks (your data cannot leak if you did not share it).
- You reduce harm to others (less third-party data, fewer identifiers).
- You reduce the risk of unfair or misleading conclusions (less “noise,” more purpose-focused input).
- You keep more control over your information.

Want to SAFE Workflow Ladder?

Stop rule:

If you cannot remove sensitive / confidential elements →
do not upload!



When you want AI help, do not start by uploading everything. Choose the safest level that still works.

- **Level 1 (Safest):** Ask a general question (no personal data).
Example: “How do I structure a CV?”
- **Level 2:** Use an anonymised summary (roles, no identifiers).
Example: “Person A and Person B had a conflict about deadlines-how can I respond?”
- **Level 3:** Use limited, de-identified data only if necessary.
Example: “Here are totals by category (no names). Help me interpret trends.”

Want to Make a risky request safe (before → after)

Rule

Share the pattern and purpose, not personal identifiers.

Prompt before:

"Here is my full CV. Improve it." (you upload a document with your full CV with address, phone number, and reference contacts etc.)

What is risky: unnecessary identifiers, third-party contact details, data that could be copied or leaked.

Prompt After (safer):

"Please improve the wording and structure of this CV. I removed personal details. Focus on clarity, achievements, and concise phrasing."

Output safety: your AI results can also leak information

Even if you share carefully, the output can still contain sensitive details if:

- Your input included private information, or,
- The AI repeats it in a summary, checklist, or draft message

Safer habits:

- Treat outputs like documents: do not paste them into public chats/groups.
- Remove names, identifiers, and unique details before sharing.
- If the output is about a person, keep it private or rewrite it more generally.
- Store sensitive outputs securely (or delete when not needed).



Want to Main activity: Turn unsafe AI use into safe AI use (10 minutes)

You will practise the full skill using fictional examples. Your goal is to produce safer requests that still solve the problem.

For each request:

- Identify unsafe parts (Never Share / Consent Gate / Fairness risk)
- Rewrite safely (minimise, anonymise, abstract)
- Add one FDC line: “What might be missing or biased?”
- Choose the workflow: Level 1 / Level 2 / Level 3 / Stop

Output:

- A short “Safe Data Playbook” you can use in daily AI use.



Want to Main activity: Turn unsafe AI use into safe AI use (10 minutes)

Practice 1:

Chat screenshot
(privacy + consent)

“Here is a screenshot of a chat with my friend [names visible]. Tell me who is right and what I should reply.”

Step-by-step tasks:

- **NSL (Never Share):** What should never be uploaded here?
- **Consent Gate:** Do you have permission to share this data?
- **Safe rewrite (Level 1 or 2):** Ask for advice without the screenshot.
- **Fairness line:** What might be missing from the story?

..... FDC line:



Want to Main activity: Turn unsafe AI use into safe AI use (10 minutes)

Practice 2:

List of people
(confidential +
fairness)

“ Please analyse this volunteer list [with names, phone numbers, and attendance problems]. Rank who is reliable.”

Why it is risky:

- Personal identifiers + third-party data
- Harmful outcome (ranking people)
- High fairness and impact risk
- **Your safer goal: make the question about Improving the system, not judging individuals.**

05



Some Data
(15 min)



Data & Credential Stealing - There is a growing trend of illegal use of data using AI



“7% increase in ransomware or other destructive attacks. 23% increase in credential theft attempts.”

Microsoft Threat Intelligence, “Microsoft Digital Defense Report 2025: Lighting the Path to a Secure Future” (Microsoft, 2025);
<https://www.microsoft.com/en-us/security/security-insider/threat-landscape/microsoft-digital-defense-report-2025>

“Data exfiltration volumes for 10 major ransomware families increased 92.7%”

Zscaler ThreatLabz, “Zscaler ThreatLabz 2025 Ransomware Report” (Zscaler, 2025);
<https://threatlabz.zscaler.com/>

Data & Credential Stealing - There is a growing trend of illegal use of data using AI



“Ransomware attacks against industrial organizations increased 87% over 2024”.

Dragos, Dragos's 8th Annual OT Cybersecurity Year in Review Is Now Available (2025);
<https://www.dragos.com/blog/dragos-8th-annual-ot-cybersecurity-year-in-review-is-now-available>.

Data & Credential Stealing - There are indications that AI systems can meaningfully assist attackers



“The actor [...] relied heavily on Claude for [malware] implementation”.

A. Moix, K. Lebedev, J. Klein, “Threat Intelligence Report: August 2025” (Anthropic, 2025);
<https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf>

“Ransomware operators APT INC deployed a likely LLM-authored data destruction script”.

CrowdStrike, “CrowdStrike 2025 Global Threat Report” (CrowdStrike, 2025);
<https://www.crowdstrike.com/en-gb/global-threat-report/>

Data & Credential Stealing - There are indications that AI systems can meaningfully assist attackers



“[Google Threat Intelligence Group] discovered a code family that employed AI capabilities mid-execution to dynamically alter the malware’s behavior. [...] Attackers are moving beyond [...] using AI tools for technical support”.

Google Threat Intelligence Group, “GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools”
(Google Threat Intelligence, 2025); <https://services.google.com/fh/files/misc/advances-in-threat-actor-usage-of-ai-tools-en.pdf>

Data & Credential Stealing



The contribution of AI systems to the trend appears to be limited and is likely secondary to other factors. However, some malicious actors would be unlikely to launch their attacks without AI systems.

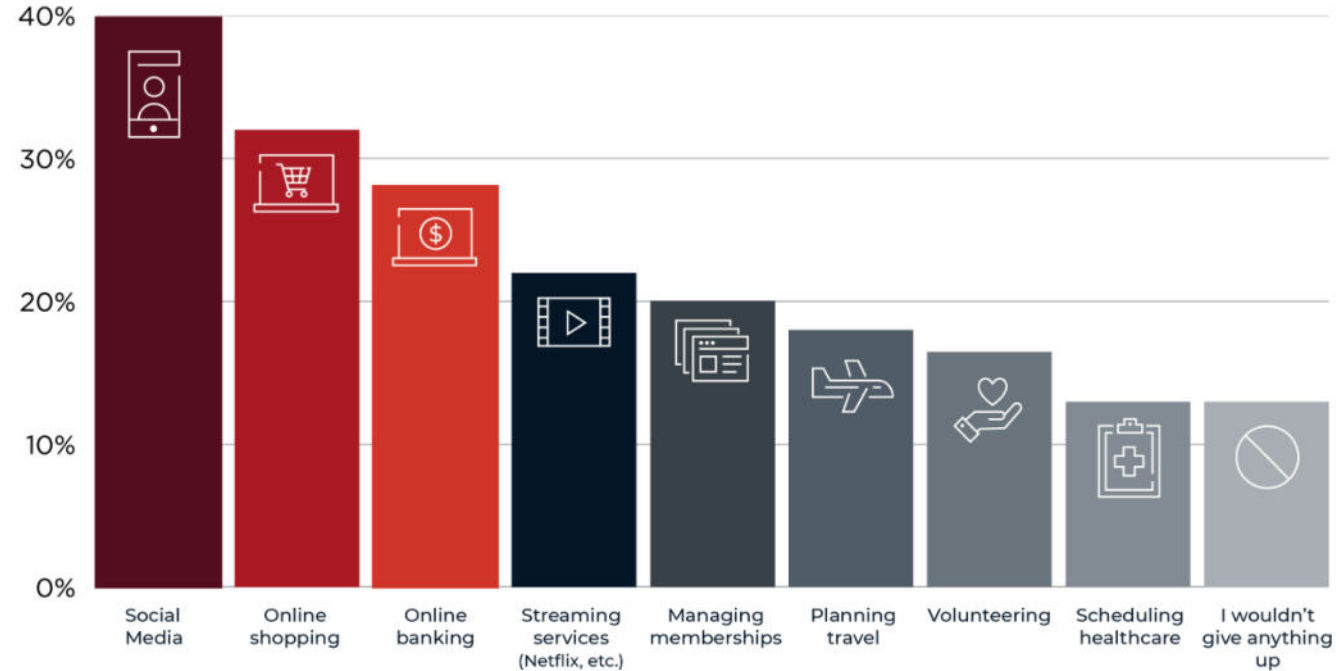
International AI Safety Report 2026 (DSIT 2026/001, 2026); <https://internationalaisafetyreport.org>

Data & Credential Stealing

Ping Identity. (2024). *Bridging the trust gap in the age of AI: 2025 consumer survey*. Ping Identity.
<https://www.pingidentity.com/en/lp/2025-consumer-survey.html>

Which types of online services do you trust least with your identity data?

Remember that AI is currently widely used and often mediates services and transactions.



Data & Credential Stealing

Ping Identity. (2024). Bridging the trust gap in the age of AI: 2025 consumer survey. Ping Identity.
<https://www.pingidentity.com/en/lp/2025-consumer-survey.html>

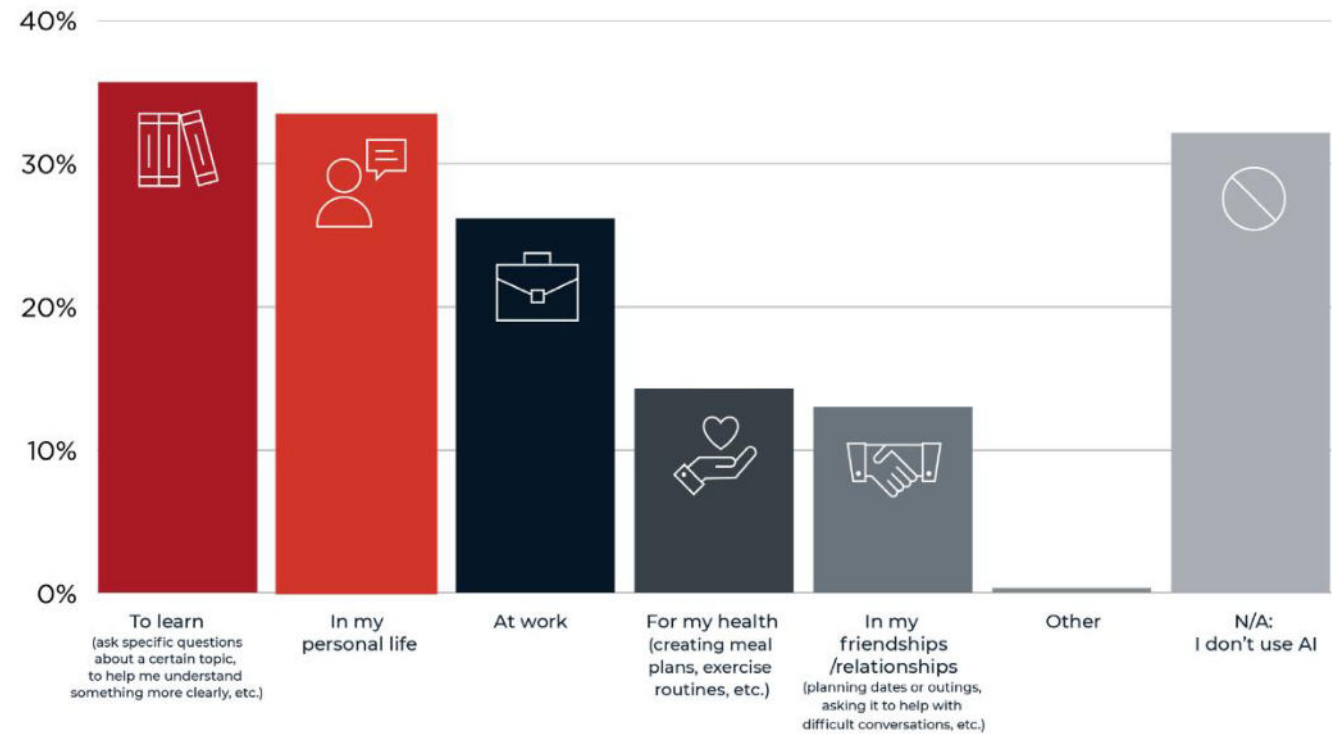
How are you currently using artificial intelligence (AI)?

“While knowledge is power, the increased use of AI in our everyday lives has only informed many consumers of its capabilities. They see firsthand how powerful this technology can be when applied to their own use cases. They’re also seeing more news stories and data points about increases in security threats and cyberattacks. Combined, it’s understandable that this contributes to a crisis of confidence in brands’ abilities to keep consumers’ identity data secure.”

Data & Credential Stealing

Ping Identity. (2024). Bridging the trust gap in the age of AI: 2025 consumer survey. Ping Identity.
<https://www.pingidentity.com/en/lp/2025-consumer-survey.html>

How are you currently using artificial intelligence (AI)?



Data & Credential Stealing - What are your concerns about AI with regard to identity security, if any?

Invasion of my personal privacy by AI programs	27%
AI-generated phishing	27%
AI being used to falsely impersonate me or a brand/person that I am engaging with	26%
Lack of transparency in how AI systems are using and storing my personal information	25%
Increased cybersecurity risks – threatening the safety of my personal data	25%
AI-generated voice cloning	24%

Continued on the next slide



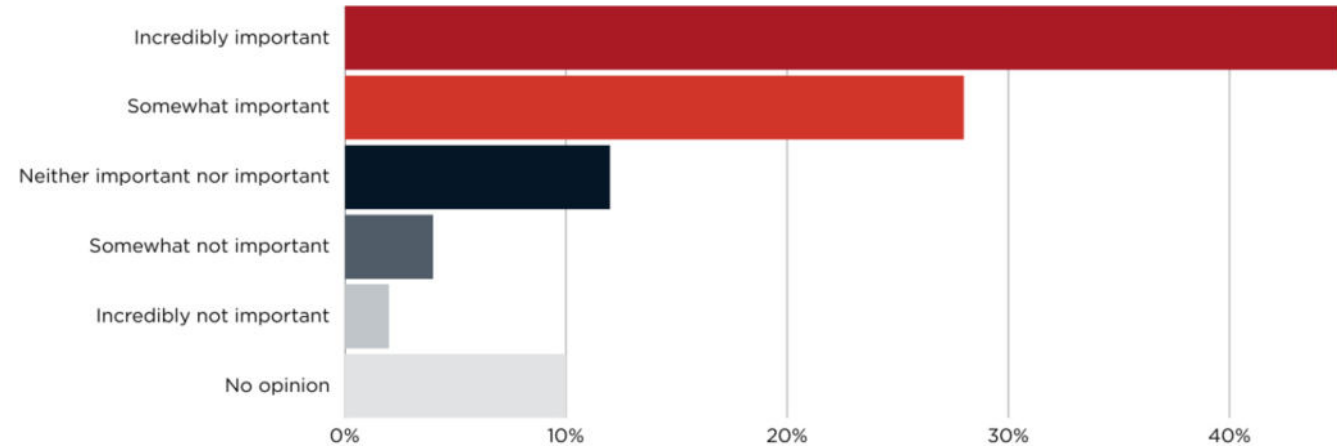
Data & Credential Stealing - What are your concerns about AI with regard to identity security, if any?

Deepfake impersonations of people I know and trust	22%
AI impersonation via chatbots	20%
Lack of accuracy/bias in AI systems	18%
I don't understand what my legal protections are for how AI uses my information	18%
Authorizing AI to work on my behalf	17%
I don't have any concerns related to AI and identity security	9%
Don't know	13%

Data & Credential Stealing

Ping Identity. (2024). Bridging the trust gap in the age of AI: 2025 consumer survey. Ping Identity.
<https://www.pingidentity.com/en/lp/2025-consumer-survey.html>

How are you currently using artificial intelligence (AI)?

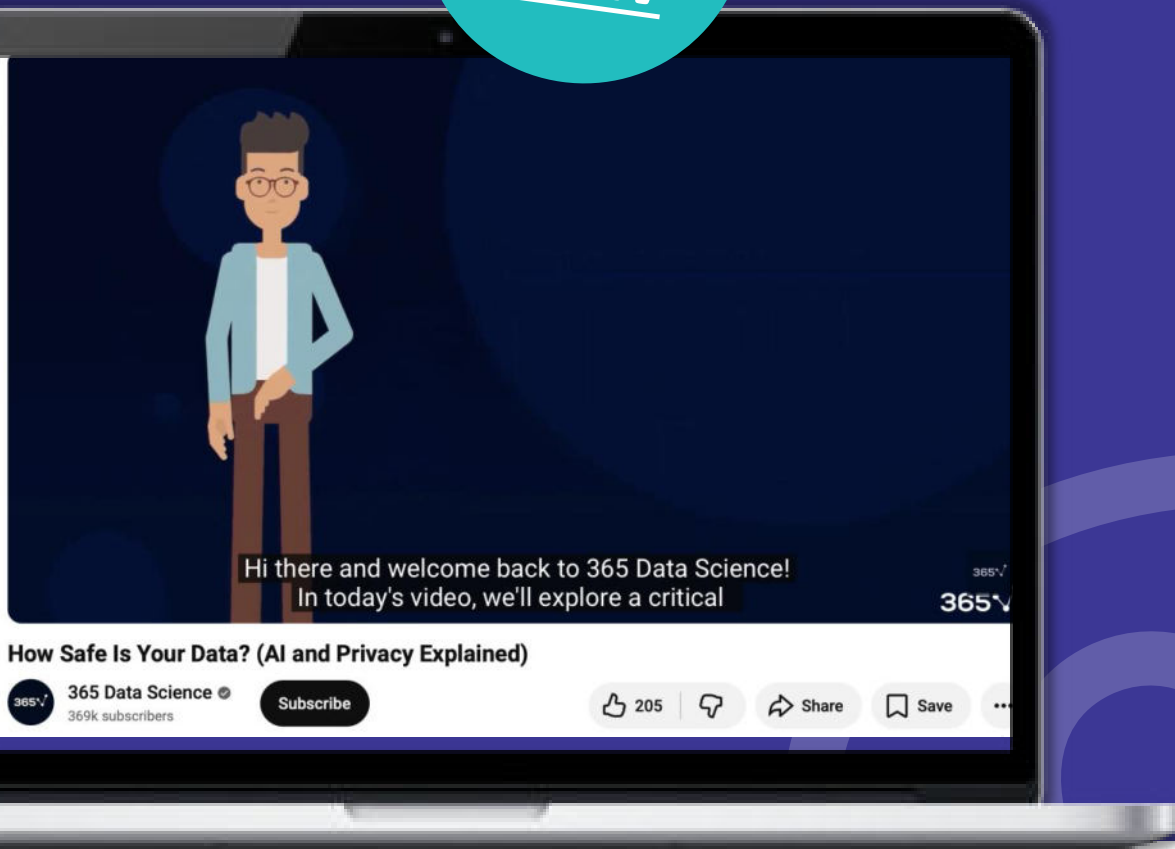


73%

of respondents reported feeling that government regulation of AI to protect their identity data is important.

Data & Credential Stealing

CLICK TO
VIEW



In today's digital age, protecting your data has never been more essential.

This video unpacks the complex interactions between artificial intelligence and the data we share online every day, from social media posts to online shopping habits.



www.valuesai.eu

Follow Our Journey



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

MODULE 06

AI and Digital Footprint

Controlling What You Leave Behind



Contents

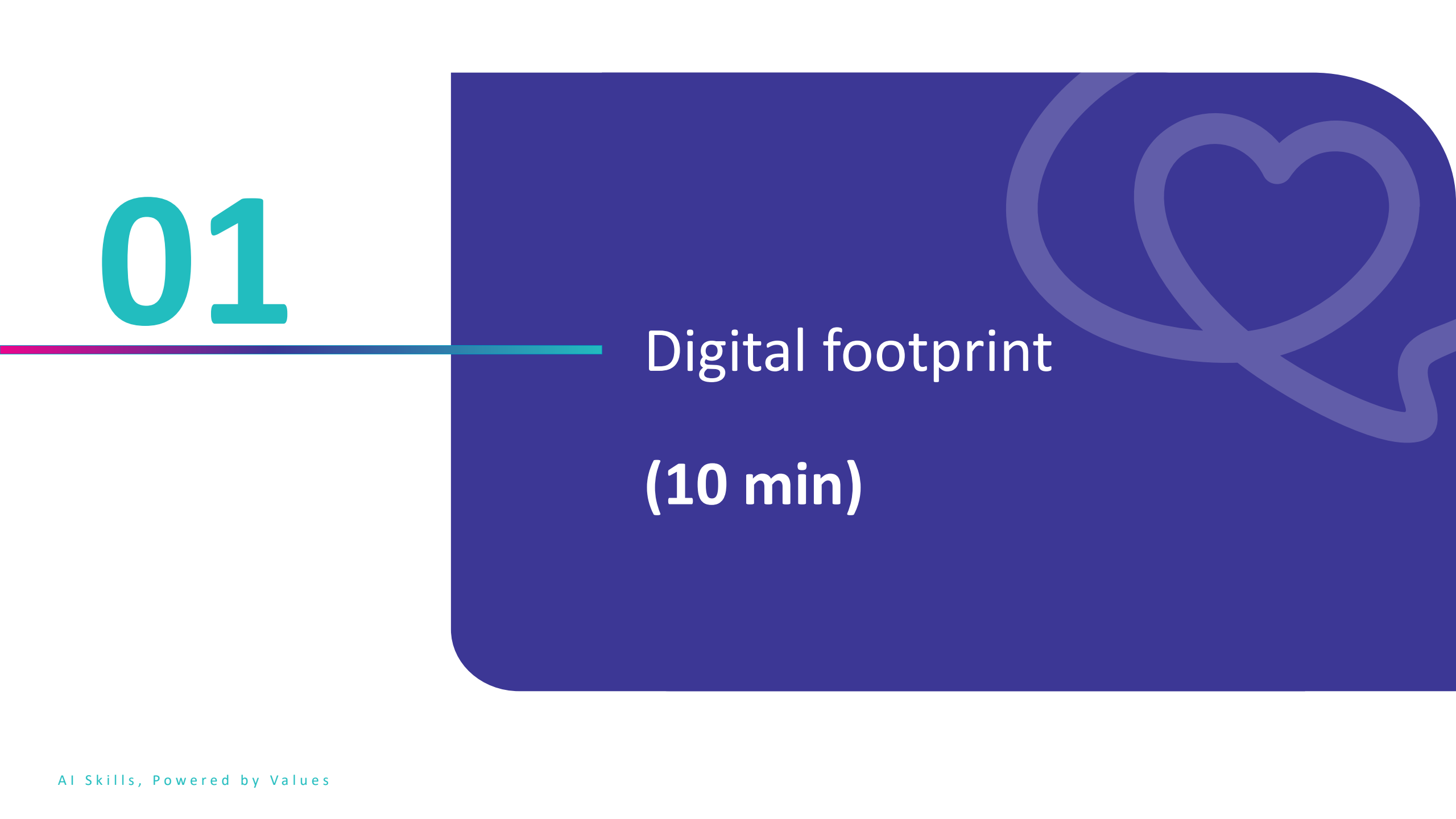
01	Digital footprint (10 min)	3
02	Types of footprints (15 min)	8
03	The “ripple effect”: consequences for you and others (25 min)	15
04	My choice: a simple checklist + my action plan (15 min)	25
05	How to protect your digital footprint (10 min)	35
06	Audit, quiz and test (20 min)	39



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

01



Digital footprint (10 min)

What digital footprint is?

A digital footprint is the collection of data about a person or organization that is left online as a result of their online activity. It includes social media profiles, posts, comments, browsing history, search queries, and information automatically recorded by websites and apps (e.g., cookies, IP address, location data).

Why does your digital footprint matter?

Your digital footprint can influence:



Privacy



who knows what
about you



Reputation



how you are
perceived



Safety



what can
be misused



Other People



your posts can
expose them too

Why does your digital footprint matter for AI

Ask yourself:

Where do I notice “algorithmic decisions” in my daily apps?



AI connects to your digital footprint because:

- AI often uses behavioural data (clicks, viewing time, likes, searches) to predict what you want next.
- Your passive footprint (permissions, location signals, device data) can influence what systems infer about you.
- AI can affect visibility and reach (what content is promoted, recommended, or hidden).
- Your choices also affect other people’s data (tagging, sharing images, screenshots).

Using AI in line with values

Remember:

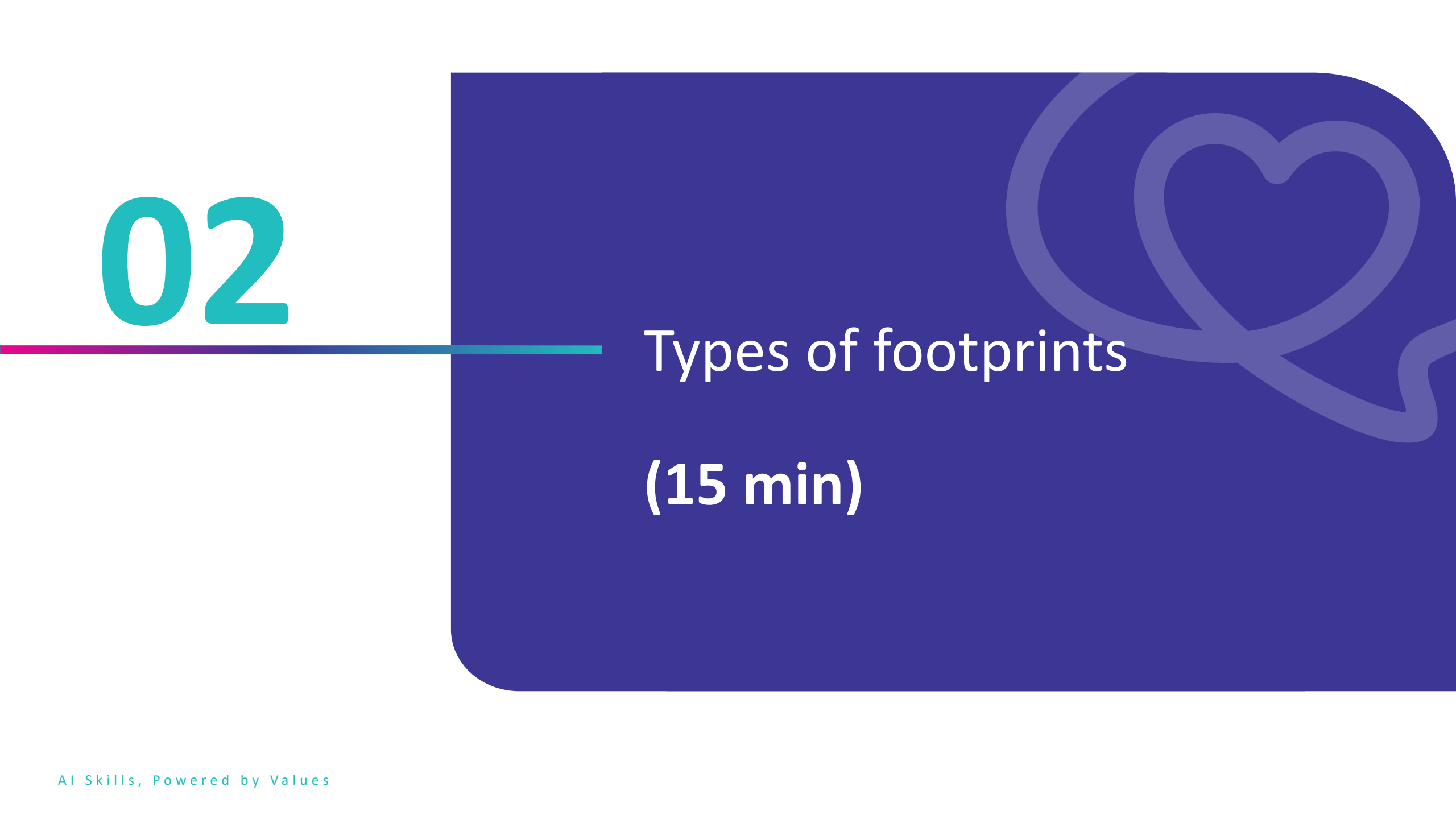
Ethical AI starts with everyday digital choices.



Four value questions to guide digital decisions (especially with AI around):

- **Agency:** Am I choosing deliberately, or just clicking automatically?
- **Privacy:** Am I sharing more data than needed (about me or others)?
- **Fairness:** Could this harm or disadvantage someone (now or later)?
- **Transparency:** Do I understand who can see it and how it can spread?

02

The slide features a large teal number '02' on the left. A horizontal line with a rainbow gradient extends from the left edge of the slide, passing behind the text 'Types of footprints'. The right side of the slide is a dark blue rounded rectangle containing the text 'Types of footprints' and '(15 min)'. In the top right corner of this blue area is a faint, stylized heart shape composed of overlapping loops.

Types of footprints

(15 min)

Active digital footprint

An active digital footprint consists of data you intentionally leave when using the internet. Examples:

- Posting messages, comments, photos,
- Sharing information in forms and surveys,
- Creating and filling out profiles on websites, companies, or platforms

With this type of footprint you have relatively more control over what you share.

Passive digital footprint

A passive digital footprint is data collected automatically, often without the user's full awareness. Examples:

- Browsing history and server logs,
- Information stored by cookies and trackers,
- Device, system, and location data (e.G., IP address),
- Interactions recorded by platforms (clicks, visit time, sessions).

This footprint is usually gathered by analytics, advertising, and tracking systems, and it is harder for users to delete or fully control.

Together, active and passive digital footprints form a person's overall digital footprint, and managing them (clearing history, limiting data, adjusting privacy settings) strongly affects online safety and privacy.



Quick activity: Sort the footprints

**Task (2 minutes)
and Select one:**

Active

Passive

Both?



- Tagging a friend in a photo
- App asking for location “always”
- Liking a public post
- Your device model in a photo’s metadata
- Posting a story with a location sticker
- Cookies tracking what you click
- Screenshotting a private chat
- A platform suggesting “people you may know”

Key Takeaway: Most people underestimate their “passive” digital footprint.

When you think about your digital footprint, you might mainly think about posts, photos, and comments. But a large part of your footprint is created without you actively posting anything, through your settings, device signals, permissions, and browsing behaviour.



Where AI meets your footprint: everyday examples

You do not need to “build or actively use AI” to be affected by it. Many online services apply AI to personalize content, rank information, and automate decisions - often using footprint signals.



Common AI-driven situations using footprint signals:

- **Ranking:** what appears first in search or in your feed
- **Targeting:** personalised ads and “suggested content”
- **Filtering & moderation:** content flagged, reduced reach, or removed
- **Pattern detection:** unusual login alerts or “suspicious activity” prompts
- **Recommendations:** “For you” feeds, suggested videos, music, or products

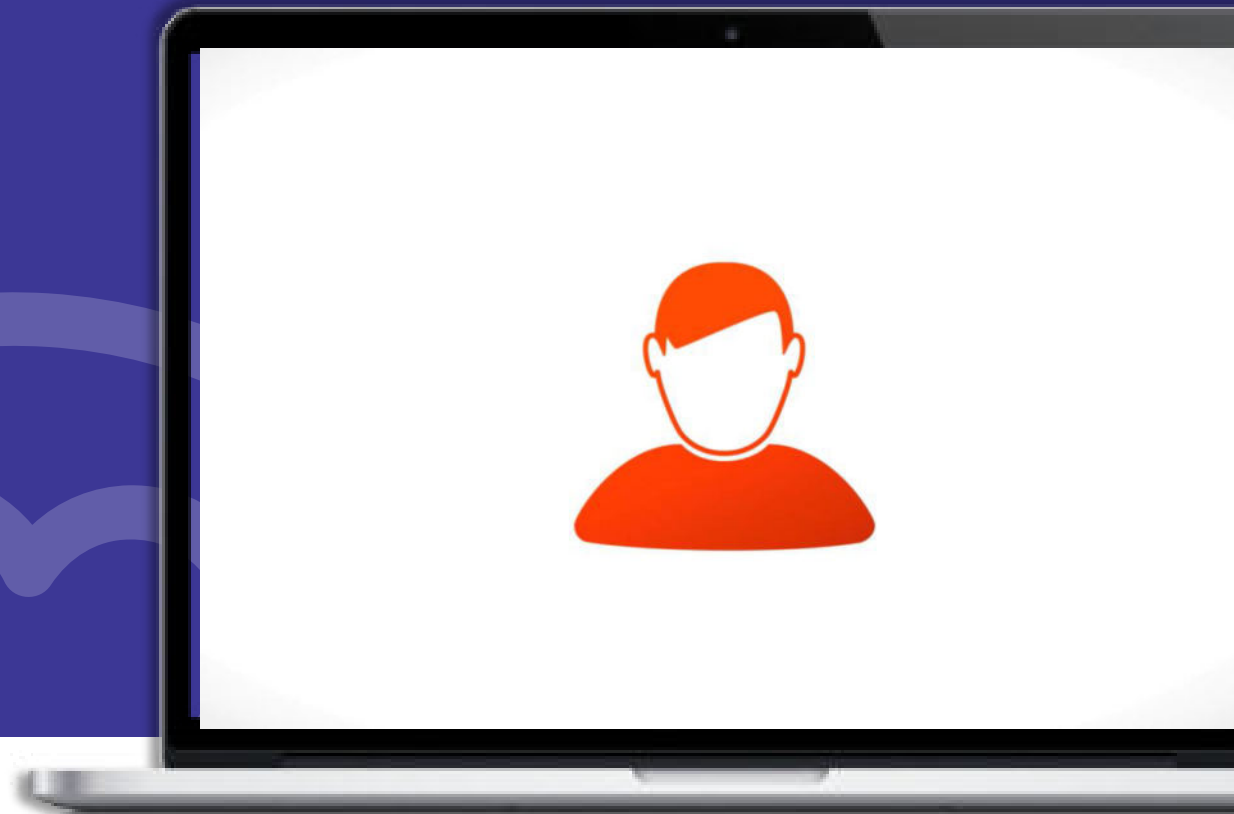
03

The “ripple effect”:
consequences for you
and others

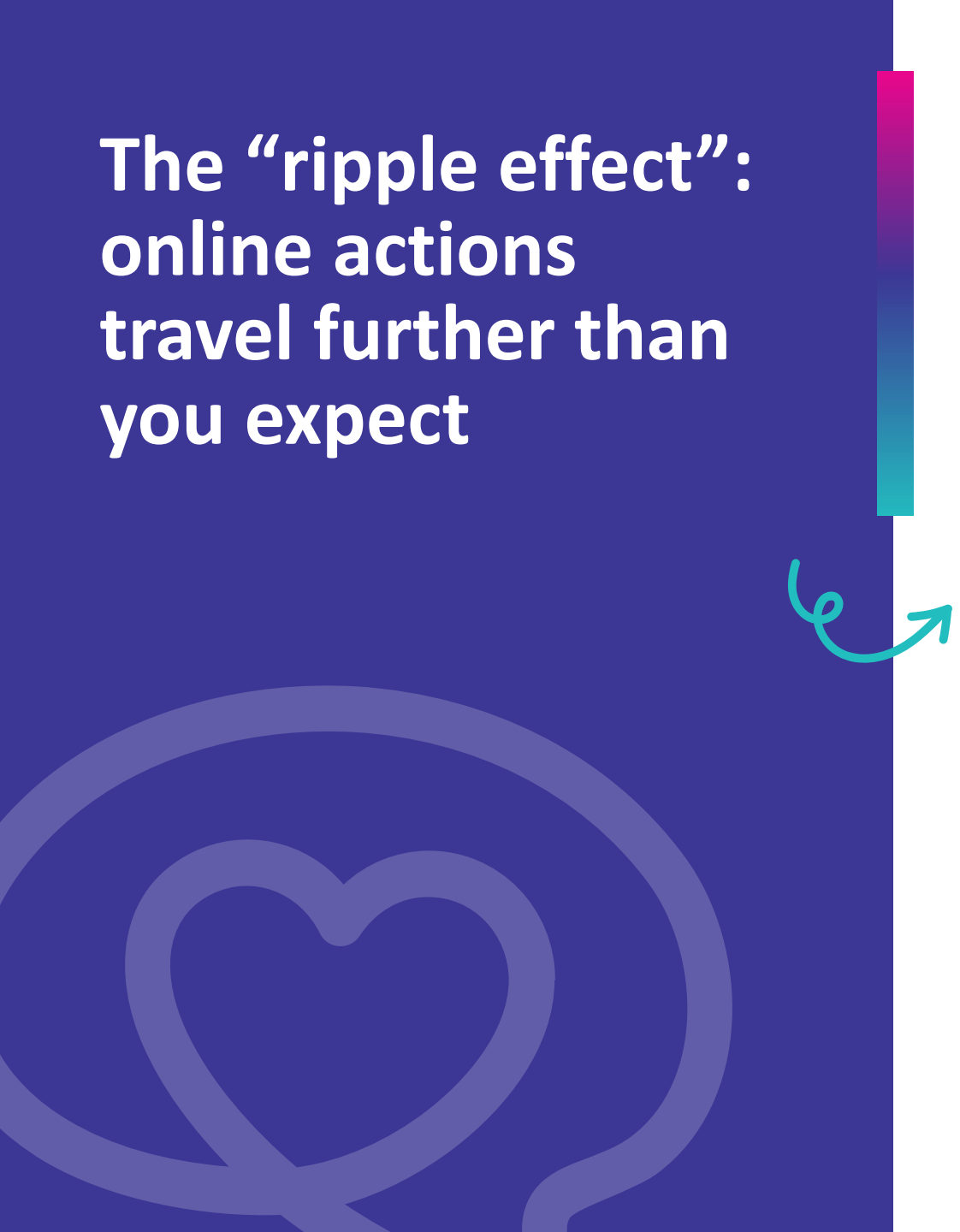
(25 min)

The “ripple effect”: online actions travel further than you expect

Online content can move quickly across people and platforms. Even small actions (a comment, a tag, a screenshot) can spread, lose context, and remain accessible longer than intended.



The “ripple effect”: online actions travel further than you expect

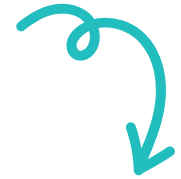


One post can lead to:

- **Copies:** screenshots, forwards, downloads
- **Context loss:** jokes or emotions get misunderstood without explanation
- **Long lifetime:** old content can resurface months or years later
- **Wider audience:** friends-of-friends, groups, or public sharing
- **Impact on others:** people visible in your content may be affected too (probably without their consent)

Consequences for you: privacy, reputation, safety, and wellbeing

A digital footprint is not automatically “good” or “bad”. It depends on context and choices. However, certain patterns (oversharing, open settings, impulsive posting) can create avoidable risks for your privacy, reputation, and even your safety.



Privacy (who can learn what about you)	Reputation (how you are perceived)	Safety (what can be misused)	Wellbeing (stress and conflict)
Public profiles, tags, location and permissions can reveal more than intended.	Posts may be seen without context; first impressions can form quickly and stick.	Overshared details can enable scams, account takeovers, or unwanted contact (not limited to your digital life).	Arguments, harassment, or regret after posting can affect confidence and mood.

Consequences for **YOU:** privacy, reputation, safety, and wellbeing



Example:

A photo with visible location + open profile settings can reveal where you are, who you are with, and your routine - without you explicitly writing it.

Ask yourself: Which risk is most relevant to me right now: privacy, reputation, safety, or wellbeing? Why?

Small changes (audience settings, permissions, pausing before posting) can significantly reduce these risks.

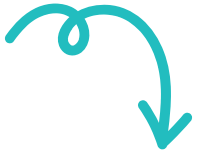
Consequences for OTHERS: respect, consent, and fairness

Common ways we expose others online:



- Posting or sharing content where others appear
- Tagging/mentioning someone (linking them to content)
- Sharing screenshots of private chats
- Revealing someone's location, schedule, or personal story
- Sharing content that could cause conflict later

Consequences for OTHERS: respect, consent, and fairness



Concrete Example

A “funny” screenshot from a private chat can be forwarded, misunderstood, and used against someone - because the original context disappears

Mini Self-Check

Ask yourself: Would I be okay if someone shared the same type of content about me, without asking?

Protecting privacy includes other people’s privacy:
consent and respect are part of acting responsibly online

Mini-scenarios: choose the most responsible option

Real-life digital decisions are often “quick moments.” This exercise helps you practise choosing an option that protects privacy, reduces harm, and respects other people, especially when content can be copied or reshared.



Scenario: A Private screenshot

A friend sends you a screenshot of a private chat and says: “This is hilarious - send it to others.”

Option 1: Forward it immediately (it is funny).

Option 2: Keep it private and ask what your friend wants to achieve by sharing it.

Option 3: Do not share it unless the people involved agree, and if in doubt, do not share.

Mini-scenarios: choose the most responsible option

Real-life digital decisions are often “quick moments.” This exercise helps you practise choosing an option that protects privacy, reduces harm, and respects other people, especially when content can be copied or reshared.



Scenario: A Private screenshot

A friend sends you a screenshot of a private chat and says: „This is hilarious - send it to others.”

Your task (30–60 seconds): Choose 1 / 2 / 3
Write one sentence:

“I chose.....because.....”

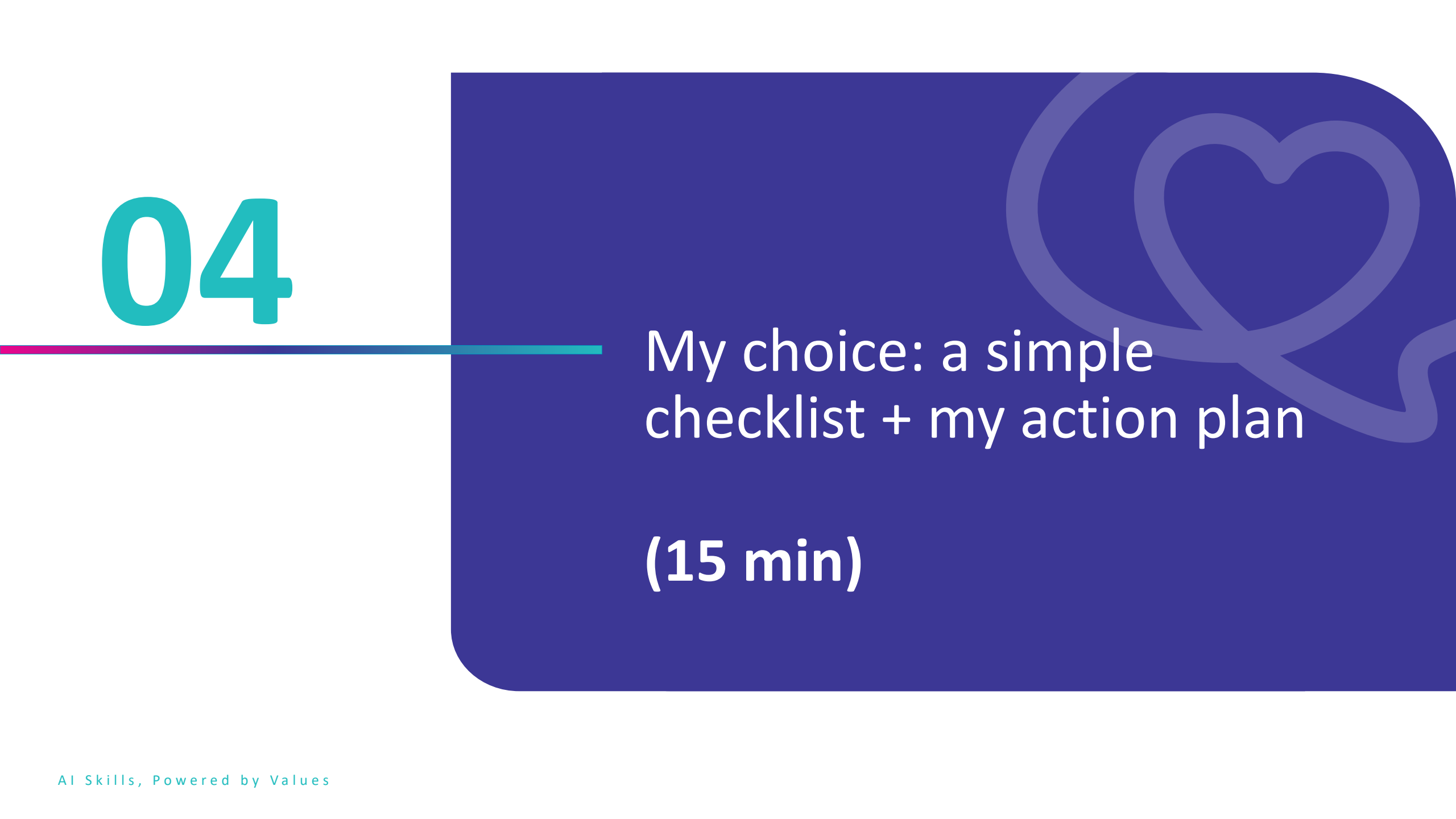
AI & Privacy - Example

What the example about:

- AI systems analyze data from everyday activities (e.g. social media, browsing, online shopping)
- Personal data becomes input for training and improving AI models
- Even simple online behaviors can be combined into detailed user profiles



04

A large, faint heart shape is visible in the background on the right side of the slide. A horizontal line with a gradient from pink to blue extends from the left edge of the slide, passing behind the number '04' and ending at the start of the dark blue box.

My choice: a simple
checklist + my action plan

(15 min)

My choice tool: STOP-THINK- CHECK-CHOOSE

This checklist is a simple habit for safer, fairer online decisions. Use it before you post, share, tag, forward a screenshot, or accept app permissions, especially when you feel rushed or emotional.



STOP	THINK	CHECK	CHOOSE
(pause for 5 seconds):	(purpose + audience):	(privacy + people + data):	(best next step):
Am I posting fast, bored, angry, excited, or under pressure?	What is the goal (inform, joke, show off, ask for help)? Who will see it now and who might see it later?	Is the audience setting correct (public / friends / close friends / group)? Does it include other people (faces, names, messages, locations)? Does it reveal extra data (location sticker, background details, metadata)?	Post / edit / crop / blur / limit audience / ask consent / do not share.

Scenario 1. Practice: use the checklist before posting

See the next Slide



You create a passive digital footprint not only by browsing, but also by what you allow apps to access.

This scenario helps you practise choosing permissions that are necessary, and declining the ones that collect more data than needed.

Scenario 1:

You install a new app (e.g., photo editor, game, or learning app).

During setup it asks for:

- Location
- Contacts
- Microphone
- Photos

STOP

Am I clicking “Allow” automatically just to finish setup quickly?

THINK

What does the app actually need to do its main job? Could I use the app with fewer permissions?

CHECK

Location: does it need location “always,” “only while using,” or not at all? **Contacts:** is access really necessary, or only “nice to have”? **Microphone:** is it needed for a feature I will use? **Photos/Media:** can I choose „selected photos only” instead of “all photos”?

CHOOSE

Give the minimum permission needed. Prefer “only while using” over “always.” Review and change permissions later if a feature truly requires it.

Scenario 2. Practice: use the checklist when other people are involved

See the next Slide



When your content includes other people, your choice affects them as well.

This scenario helps you practise making a decision that respects privacy and consent, because sharing, tagging, and forwarding can be difficult to undo.



Scenario 2:

You have a group photo from a meeting, event, or night out. You want to post it and tag your friends. One person in the photo has previously said they do not want to be posted online.

STOP

Am I about to post quickly because I want to share the moment right now?

THINK

What is the purpose of posting? Is tagging necessary? Who will see it now - and who might see it later?

CHECK

Consent: Has everyone agreed to be shown or tagged /
Audience: Is this public, or limited to a small group /
Risk: Could this photo harm or embarrass someone later?
Alternatives: Can I crop, blur, or choose a different photo?

CHOOSE

Pick the best next step: ask the person first / post without them visible / blur or crop / limit audience / do not post.

Scenario 3. Practice: use the checklist for app permissions

See the next Slide



You create a passive digital footprint not only by browsing, but also by what you allow apps to access.

This scenario helps you practise choosing permissions that are necessary and declining the ones that collect more data than needed.

Scenario 3:

You install a new app (e.g., photo editor, game, or learning app).

During setup it asks for:

- Location
- Contacts
- Microphone
- Photos

STOP

Am I clicking “Allow” automatically just to finish setup quickly?

THINK

What does the app actually need to do its main job? Could I use the app with fewer permissions?

CHECK

Location: does it need location “always,” “only while using,” or not at all?

Contacts: is access really necessary, or only “nice to have”?

Microphone: is it needed for a feature I will use?

Photos/Media: can I choose “selected photos only” instead of “all photos”?

CHOOSE

A safer default approach: Give the minimum permission needed. Prefer “only while using” over “always.” Review and change permissions later if a feature truly requires it.

Three high-impact controls: small changes, big effect



You do not need to master every privacy setting to improve your digital footprint. These three controls have a strong impact across most platforms and apps - and they are realistic to check even if you are short on time.

Your digital footprint improves fastest when you control audience, location, and discoverability.

Three high-impact controls: small changes, big effect

Core content (self-study guidance, generic-not platform-specific) 

- **Audience visibility (who can see your content):** Check whether your default is public, friends, close friends, or private. Remember: “friends of friends” or open groups can still be a wide audience.
- **Location sharing (when your location is revealed):** Avoid sharing real-time location publicly. Prefer location access “only while using” (or switch it off if not needed). Watch for hidden location clues in photos (background, landmarks, routine places).
- **Discoverability & contact syncing (how people can find you):** Review whether your profile can be found by phone number/email. Check whether apps sync your contacts by default. Decide if you want to be searchable, or only reachable by people you choose.

05



How to protect your
digital footprint

(10 min)

○ Practical Rules



01

Limit shared information on social media



02

Tighten up privacy settings on social media



03

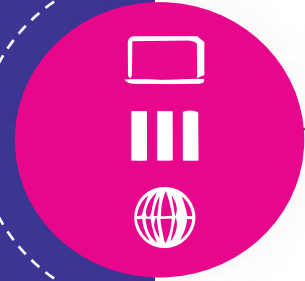
Limit the amount of data placed on the internet



05

Occasionally clean browser cookies, and other tracking files on your personal devices

○ Practical Rules



06

Use VPN services, if and when possible



07

Anonymise, disallow, or restrict location access of tracking by applications



08

Install and keep antivirus software updated on personal devices



09

Periodically review both financial/credit and medical/health information

○ Practical Rules



10

Do not open attachments or access links from unknown or questionable sources.



11

Setup credit reporting notifications



12

Continually update passwords and passwords protections, especially after being informed of a data breach

06



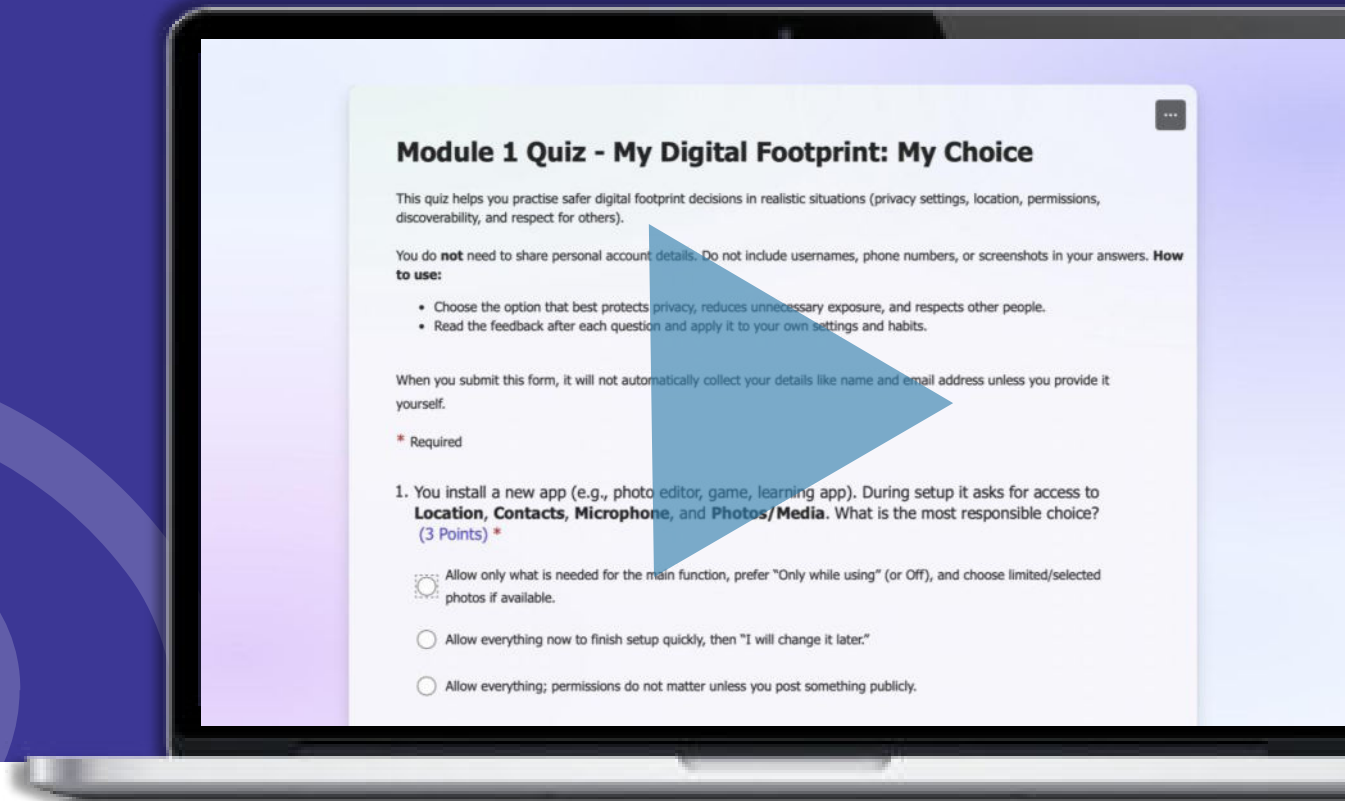
Audit, quiz
and test

(20 min)



Quiz

This quiz helps you practise safer digital footprint decisions in realistic situations (privacy settings, location, permissions, discoverability, and respect for others).



Test: Check your digital footprint

Unsure about your digital footprint? Answer “Yes” or “No” to the questions below, then calculate the total number of “Yes” answers to determine the complexity of your footprint.

1. Do you make purchases online?
2. Have you signed up for coupons by creating an account?
3. Have you registered or subscribed to newsletters or blog updates?
4. Have you downloaded and used shopping apps?
5. Have you opened a new credit card account?
6. Do you use a mobile banking app?
7. Have you bought or sold stocks?
8. Have you ever registered your email address with a gym?
9. Do you receive health care?

Test: Check your digital footprint

Unsure about your digital footprint? Answer “Yes” or “No” to the questions below, then calculate the total number of “Yes” answers to determine the complexity of your footprint.

Source: *An insider's digital footprint and associated risks*. Job aid. Defense Counterintelligence and Security Agency (CDSE), 2022
<https://www.cdse.edu/Portals/124/Documents/jobaids/insider/insiders-digital-footprint.pdf>

Y N

10. Have you ever used apps to track your activities and workouts?

11. Do you subscribe to an online publication or news source?

12. Have you ever reposted articles and information you've read?

13. Do you use social media on your computer or devices?

14. Do you interact with friends online?

15. Have you ever shared information, data, and photos with your online connections?

16. Have you ever joined a dating site or app?

Test: Check your digital footprint

RESULTS



Yes Answers <5

A digital footprint that is less complex than most internet users

Yes Answers 5 - 12

A digital footprint that is average for most internet users

Yes Answers 12+

A digital footprint that is more complex than most internet users

Digital Footprint Audit



Now you will practise recognising common footprint risks and choosing practical improvements. You can also complete this without sharing personal information by using a fictional example, or optionally by checking one setting on your own device privately.

Private own-device check - Check one privacy/permission setting on your device or account. If you are working in a group, do so without sharing your screen or details.

Digital Footprint Audit - Choose one of your Social Media Apps

The goal is not to delete your online life—it is to reduce unnecessary exposure and make your footprint intentional.

Complete these three lines:

Risk 1 → Fix 1: _____ → _____

Risk 2 → Fix 2: _____ → _____

Risk 3 → Fix 3: _____ → _____

First action (today): _____

Step 1: Identify 3 footprint risks

- Public visibility too wide (posts/profile discoverable)
- Location revealed (real-time or routine places)
- Tagging/mentions without consent
- Permissions too broad (contacts/microphone/location “always”)
- Personal details visible (school/work, phone, email, routine)

Step 2: Choose 3 Improvements (Match one to each risk)

- Limit audience / switch to private / review default visibility
- Remove or limit location sharing
- Ask consent before tagging or posting others
- Reduce permissions to „minimum necessary”
- Remove sensitive details from profile

Step 3: Decide your “first action”

- What is the one change you would make first if you had only 2 minutes

Digital Footprint Audit

A digital footprint improves through small, consistent actions - not one perfect setting. This short plan helps you turn today's audit into concrete steps you can actually complete.

Make it doable: If your plan feels too big, reduce it to one setting + one habit. Your digital footprint is built by habits. A small plan you follow is better than a big plan you forget.

1. One change I will make today (5 minutes)

Examples:

- Limit profile visibility
- Remove location sharing
- Reduce one permission
- Review discoverability

Write:

Today I will:

.....

.....

Digital Footprint Audit

A digital footprint improves through small, consistent actions - not one perfect setting. This short plan helps you turn today's audit into concrete steps you can actually complete.

Make it doable: If your plan feels too big, reduce it to one setting + one habit. Your digital footprint is built by habits. A small plan you follow is better than a big plan you forget.

2. Two changes I will make this week (10–20 minutes total)

Examples:

- Review privacy defaults
- Clean old public posts
- Check app permissions
- Update password settings (if relevant)

Write:

This week I will:

.....

.....

Digital Footprint Audit

A digital footprint improves through small, consistent actions - not one perfect setting. This short plan helps you turn today's audit into concrete steps you can actually complete.

Make it doable: If your plan feels too big, reduce it to one setting + one habit. Your digital footprint is built by habits. A small plan you follow is better than a big plan you forget.

3. One habit rule I will follow going forward (daily choice)

Examples:

- "I will use STOP-THINK-CHECK-CHOOSE before sharing"
- "I will ask consent before tagging"
- "I will avoid real-time location"

Write:

My rule is:
.....
.....



www.valuesai.eu

Follow Our Journey



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

MODULE 07

Acting Legally with AI

Rights, Consent and Responsible Use



Contents

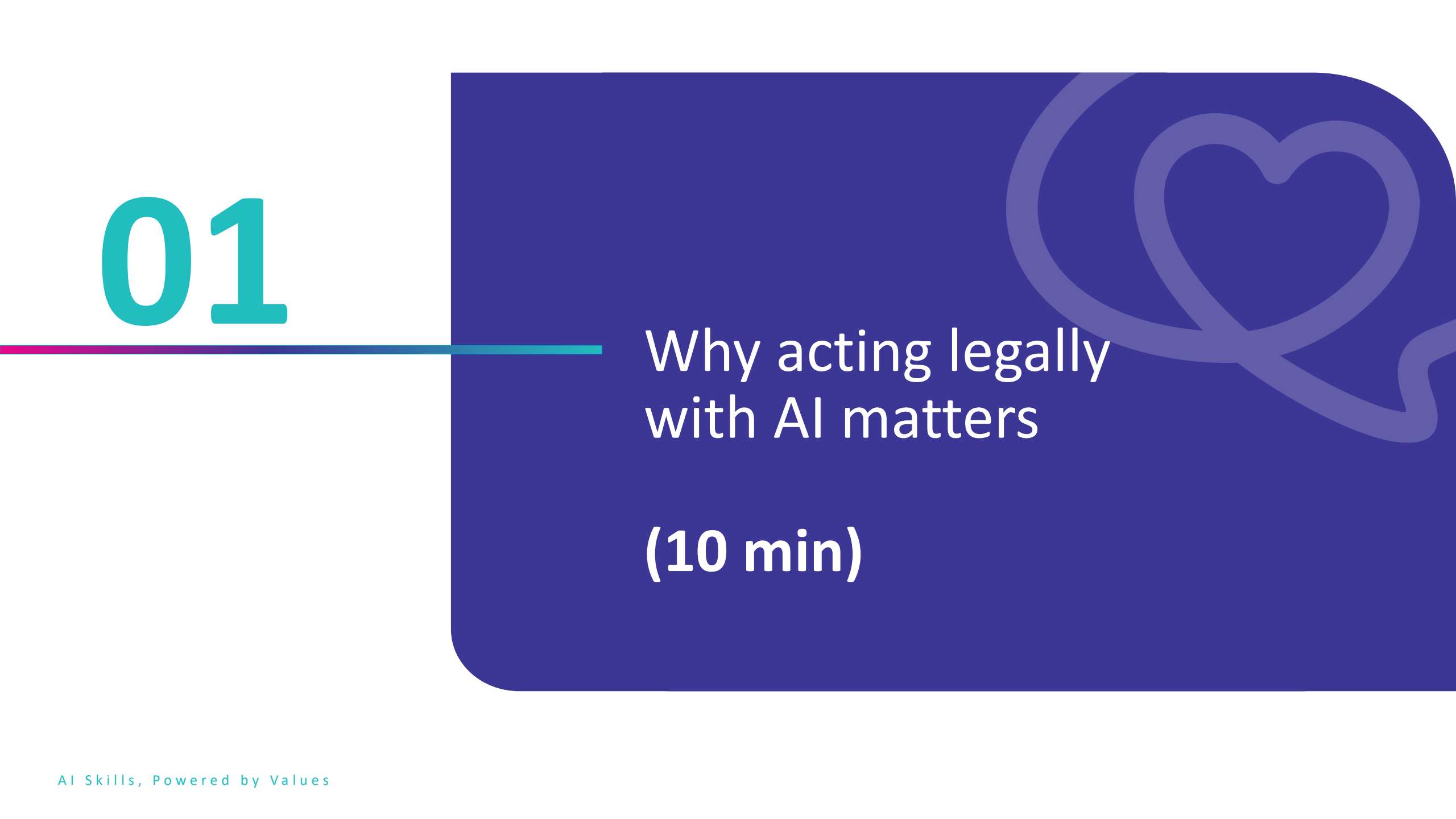
01	Why acting legally with AI matters (10 min)	3
02	Voice rights and AI (10 min)	9
03	Image and likeness rights (10 min)	14
04	Copyright in the AI era (10 min)	19
05	AI law: EU AI Act and GDPR (10 min)	23



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

01

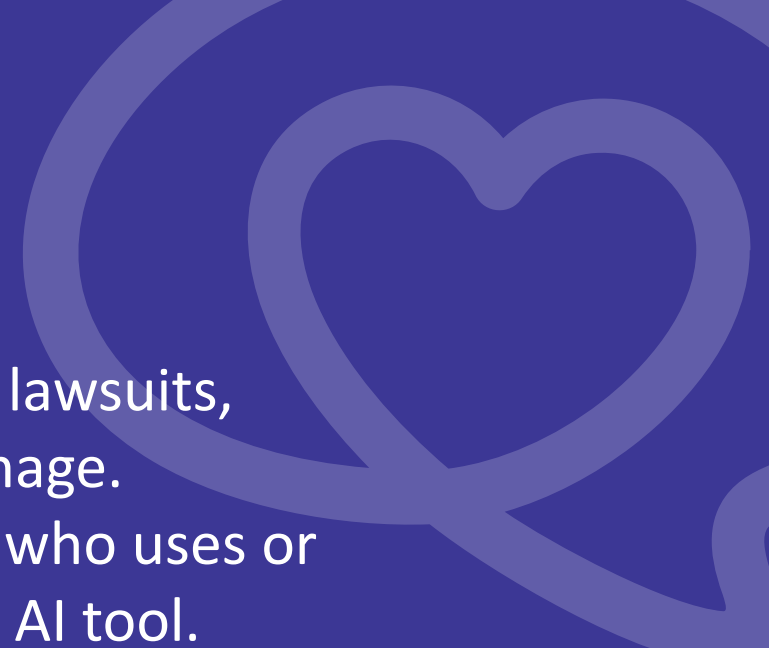
A decorative graphic featuring a large, stylized heart shape in the background. A horizontal line with a gradient from pink to blue extends from the left edge of the slide, passing behind the heart and ending at the start of the main text area.

Why acting legally
with AI matters

(10 min)

AI tools make it easy - in seconds - to clone voices, generate faces, copy styles, and reuse text. Just because something is possible does not mean it is lawful.

Why this matters:



- Legal violations can lead to fines, lawsuits, takedowns, and reputational damage.
 - Liability often falls on the person who uses or publishes the output - not on the AI tool.
 - Lawful use is also a form of respect: for people, creators, and institutions.
 - "Legal" is more than copyright - it covers personality rights, data protection, and AI-specific rules.
- 

Example: Asking AI to summarise a public news article is generally fine. Asking AI to clone a colleague's voice for a prank - without their consent - is not

AI use creates three legal risks: personality, copyright, and regulation

**When you use AI,
three legal layers
can be triggered at
the same time:**

1. **Personality rights risk:** Using someone's voice, face, or name without consent - even in fiction or satire - can violate their image / publicity rights.
2. **Copyright risk:** AI inputs and outputs may copy protected works. Generated images can resemble existing artworks; generated text can reproduce passages.
3. **Regulatory risk:** GDPR, the EU AI Act, sector laws (health, finance, education), and contracts may impose duties - especially around disclosure and personal data.



DO



Ask: Whose rights does this touch?
Could a court agree?



Use checks: Voice Gate + Image Checklist + Copyright Check



DON'T

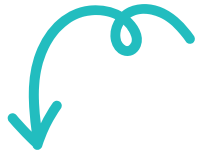


Assume AI outputs are legally clean

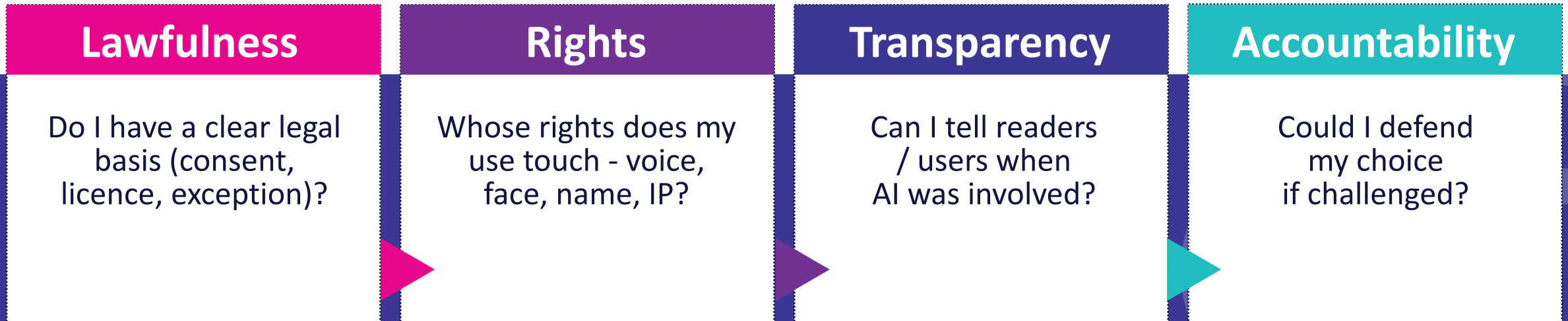


Use AI to imitate real people without permission

For Legal Use



Use four questions whenever you use AI in a way that could touch the law



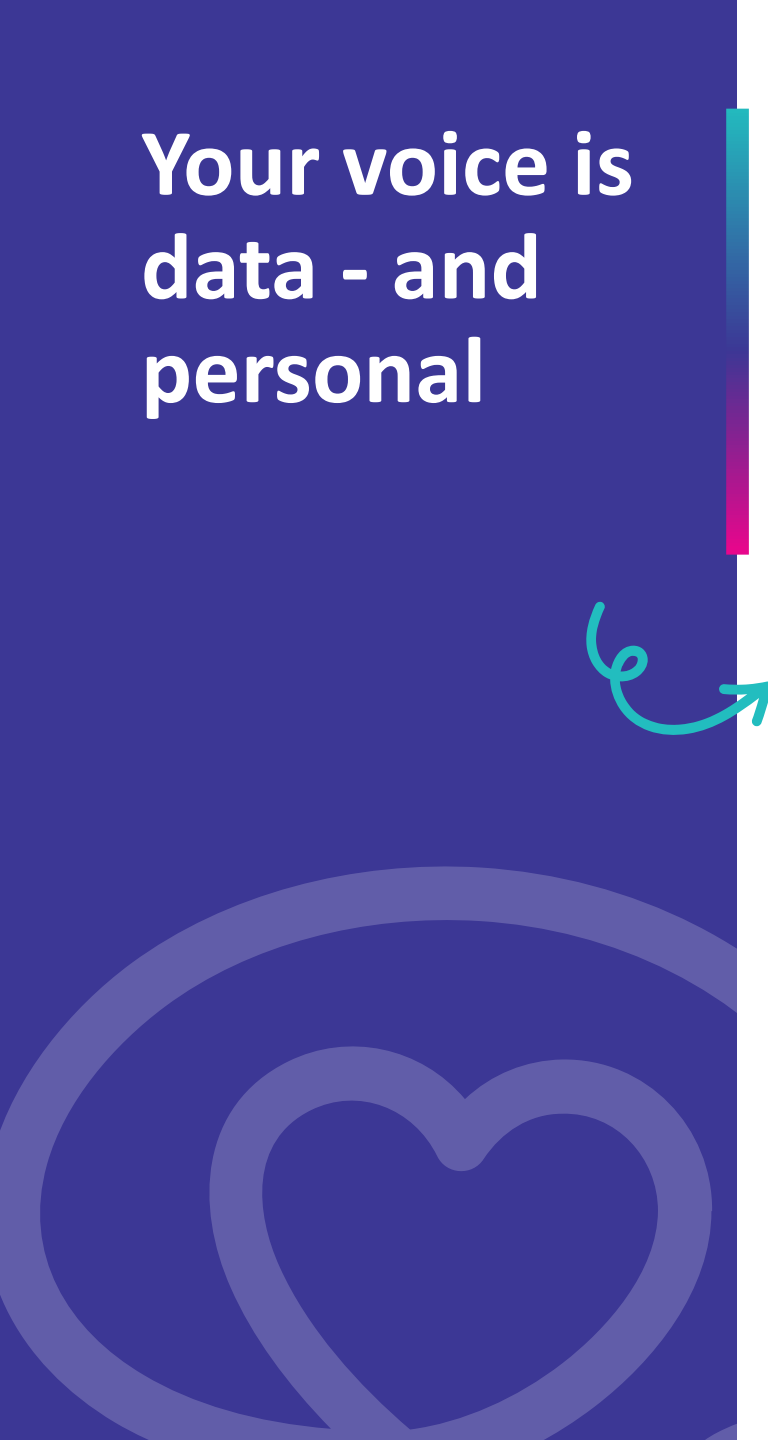
Lawful AI use = informed choices guided by values

02

A decorative graphic consisting of a horizontal line with a pink-to-blue gradient, extending from the left edge of the slide. To the right of the line, there is a large, stylized heart shape composed of concentric, hand-drawn style lines in shades of blue and purple. The heart is positioned in the upper right quadrant of the slide, partially overlapping the dark blue background area.

Voice rights and AI (10 min)

Your voice is data - and personal



**AI can clone a voice from a few seconds of audio.
That changes what “voice” means in law:**

- A voice identifies a person → under GDPR, voice can be personal data, and voiceprints can be biometric data
- A voice has commercial value - actors, narrators, singers, and public figures rely on it
- “Sounds like” is enough to harm someone (reputation, fraud, deception)
- Performers rights protect existing recordings - even short samples used to train a voice model can be problematic

Key Takeaway: Voice rights and the law

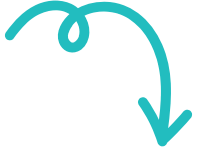


Cloning, modifying, or imitating a real voice can trigger several legal regimes at once:

- **Personality / publicity rights:** many countries protect a person's voice as part of their identity. Commercial use without consent is usually unlawful.
- **GDPR (EU):** voice processing usually requires a legal basis. Voiceprints are special-category data and need stronger justification.
- **Performers' rights:** actors and singers have rights over their recorded performances; using them to train AI may need separate clearance.
- **Defamation and fraud:** voice clones used to impersonate or deceive can lead to civil and criminal liability.
- **EU AI Act:** AI-generated audio that imitates a real person (voice deepfake) must generally be labelled.

Voice Consent Gate (3 Questions): Decide before you clone

Use this gate whenever your AI use involves a real person's voice (clone, impression, modification):



Permission:

Do I have written consent (or a clear licence) for this voice and this purpose?

Identifiability:

Could the result be confused with the real person?
Could it harm or mislead?

Disclosure:

Will I clearly label the audio as AI-generated where the law or policy requires it?

If you answer No / Not sure to any question → use a fully synthetic, non-identifiable voice or a properly licensed stock voice. "Not sure" counts as a warning sign.



● Micro-task: Voice scenarios

Task (60–90 seconds):
Mark each as Lawful /
Risky / Do not do - and
write a short reason.

1. Cloning a politician's voice for satire on
2. Social media without disclosure
3. Using a stock AI voice (commercial licence) for a video
4. Cloning your own voice for your own podcast intro
5. Cloning a colleague's voice „for fun” without telling them
6. Using a deceased actor's voice in an advertisement
7. Without estate permission
8. Generating a voice that just „sounds similar”
9. To a famous singer

03

Image and
likeness rights

(10 min)



Image rights and AI: two layers to check

AI can generate, edit, swap, and age faces in seconds. Two legal layers usually apply:



- **Right of the person depicted (image / likeness rights, GDPR for biometric data)** - you usually need consent to use a recognisable face.
- **Right of the photographer (copyright in the photo itself)** - you need a licence to use the photo, even if the person agrees.

Image rights and AI: two layers to check

**What counts as “using
someone's image”:**



- Feeding their photo into AI to generate variations
- Swapping a face into a new background or scene
- Ageing, undressing, or stylising someone's photo
- Training a custom model on their pictures

Key Takeaway: Image Use Checklist

Before using AI with images of real people, check four things:

1. **Right to the photo:** do I own it, or do I have a licence (stock, Creative Commons, etc.)?
2. **Right to the person:** do I have consent for this specific use? Old consent for one purpose does not cover new AI uses.
3. **Context:** could the new context embarrass, harm, or mislead? (sexual, political, medical, criminal contexts are especially risky)
4. **Disclosure:** do I need to label the image as AI-generated or AI-edited (EU AI Act, platform rules, journalism standards)?



If you cannot answer all four - do not generate, do not publish



● Micro-task: Image scenarios

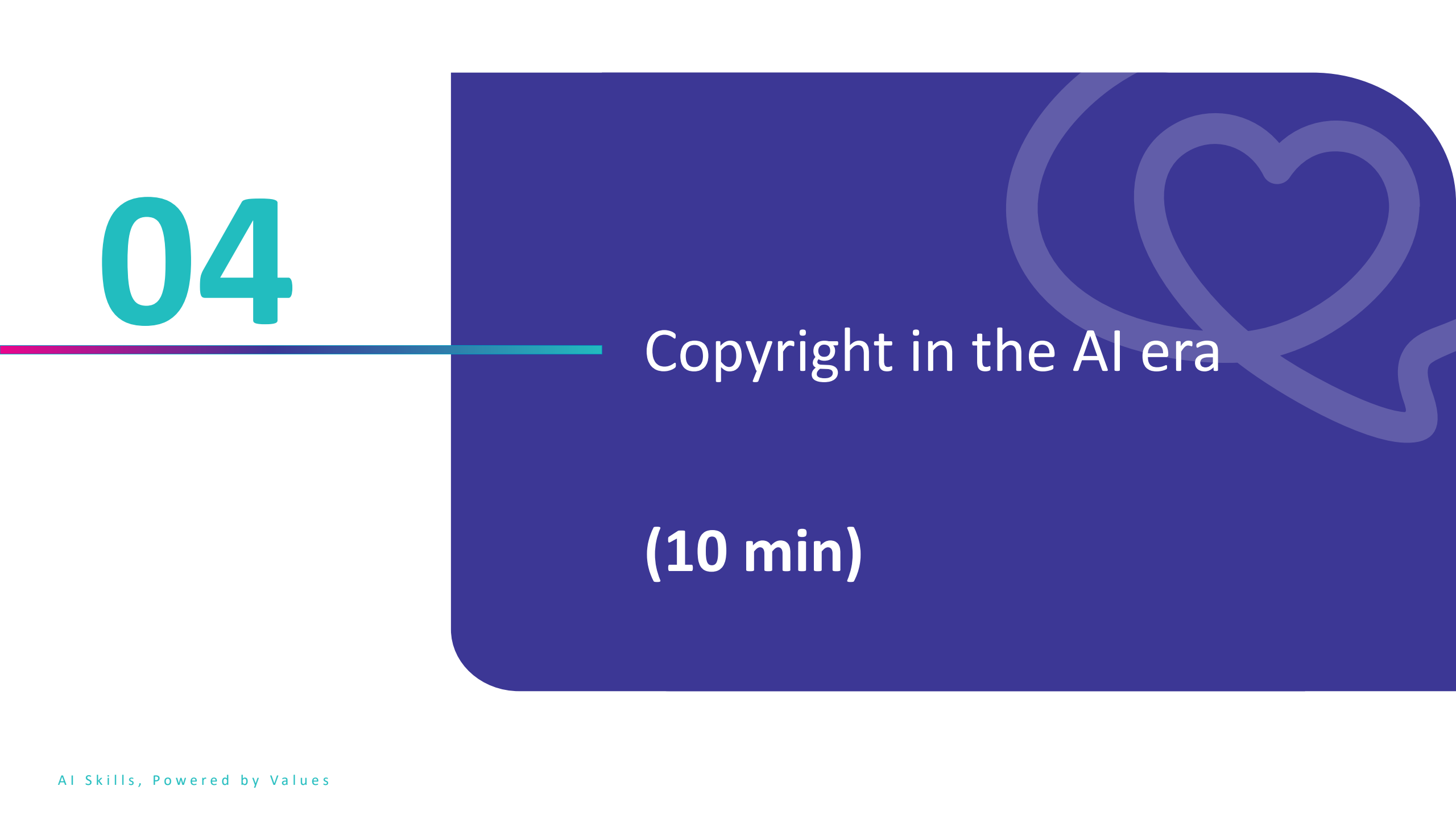
Task (60–90 seconds):

For each, decide Pass / Risky / Stop - then write one short reason

Self-check hint: If the result could mislead viewers about a real person - Risky or Stop

1. Generating a fictional avatar (no real person) for your own profile picture
2. Putting a colleague's face on a meme without telling them
3. Using AI to restore an old family photo for private use
4. Generating a fake “before/after” image of a public figure for political content
5. Using a stock AI portrait (with commercial licence) on your website - clearly labelled as AI
6. Editing your own selfie with an AI filter and posting it

04

A large, faint heart shape is visible in the background on the right side of the slide. A horizontal line with a gradient from pink to blue extends from the left edge of the slide, passing behind the title text.

Copyright in the AI era

(10 min)

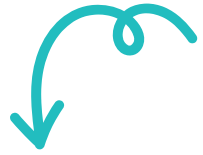
Training data and copyright

AI models are trained on huge amounts of text, images, code, and audio. Most of it is copyrighted



- Whether training is lawful is contested and varies by jurisdiction (EU has TDM exceptions with opt-out for rights holders; the US relies on fair use, which is being litigated).
- For end users, the practical risk is the output: AI can unintentionally reproduce a protected work - verbatim text, a famous logo, a recognisable artistic style.
- Style is generally not copyrightable, but trademarked characters (Mickey Mouse, Pikachu) and specific compositions are.
- “The AI did it” is not a legal defence - the person who publishes the output is usually liable.

Practical rule: Always review AI outputs for accidental copying before publishing - especially images, logos, and longer passages of text



Who owns AI-generated content?

Ownership (copyright):

- **US Copyright Office:** purely AI-generated works are not copyrightable. Only the human-authored parts (prompt selection, edits) may be protected.
- **EU:** similar approach - copyright requires a human author and a creative choice.
- **Practical effect:** your AI outputs may be free for anyone to copy. Heavy human editing strengthens your claim.

Usage rights (terms of service):

- **Tool providers** (OpenAI, Anthropic, Midjourney, Google, etc.) usually grant you a licence to use outputs commercially - but read the terms.
- **A licence is not ownership:** you can use the output, but you may not be able to stop others from using a similar one.
- **Infringement risk stays with you:** even when an AI generates a logo, you can still be sued if it copies an existing trademark.

Plagiarism, attribution, and AI

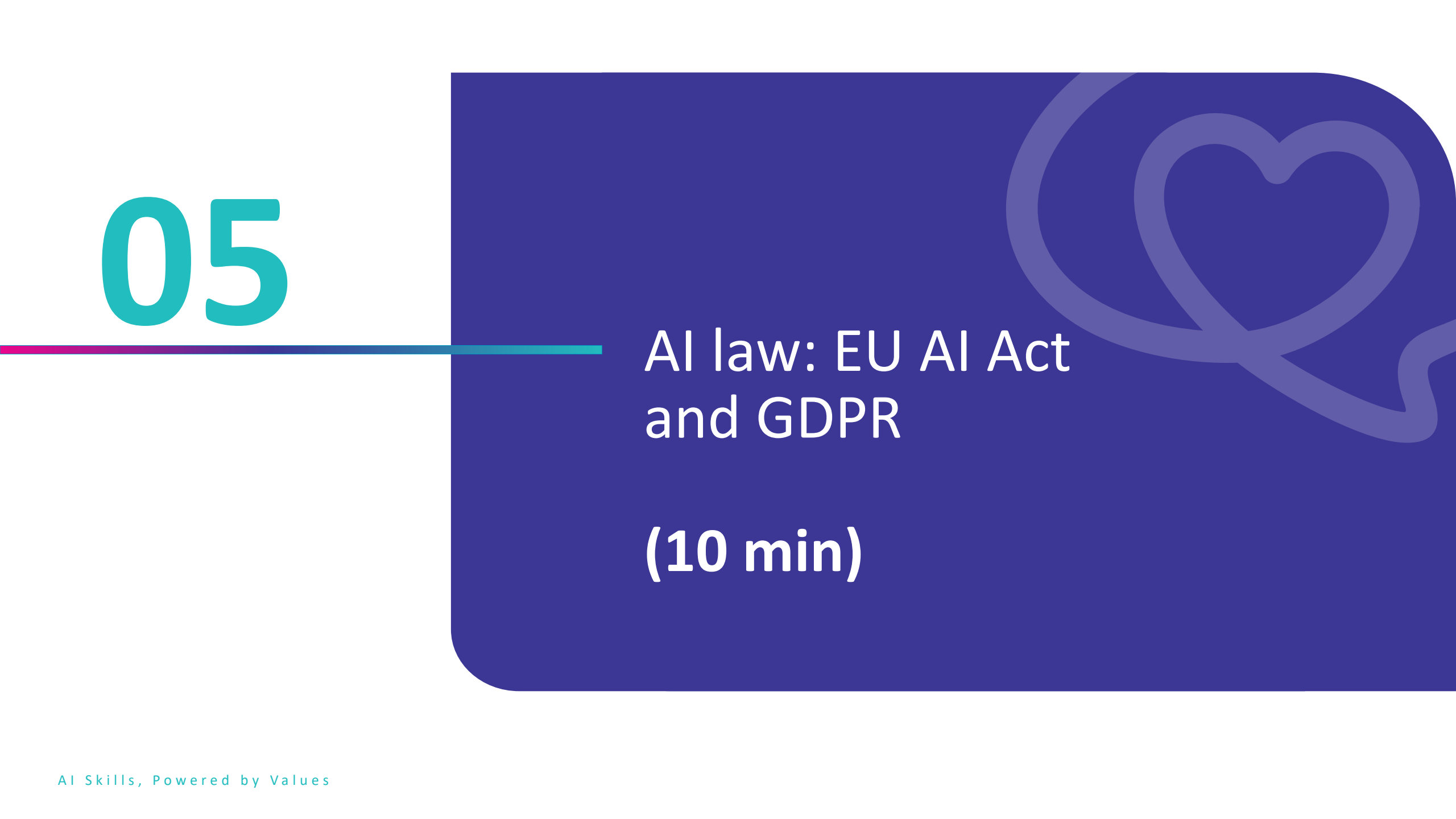
Copyright and plagiarism are different:



- **Copyright is law** - about who can copy or reuse a work.
- **Plagiarism is an ethics/academic rule** - about taking credit for work that is not yours, regardless of permission.
- **Universities, journals, and many employers** now require disclosure of AI use; some prohibit it for graded work.
- **Common rules emerging:** AI cannot be a co-author; you must declare which parts were AI-assisted; you remain responsible for accuracy.

Best practice: keep your prompts and drafts; cite the tools you used; verify every fact, quote, and citation - AI is known to fabricate them.

05

A decorative graphic on the right side of the slide. It features a large, stylized heart shape composed of concentric, slightly irregular lines in shades of blue and purple. A horizontal line, transitioning from pink to blue, extends from the left edge of the slide and ends at the heart graphic.

AI law: EU AI Act
and GDPR

(10 min)

EU AI Act in brief: a risk-based law

The AI Act sorts AI uses into four risk levels. Your obligations depend on the level

Prohibited	High-Risk	Limited-risk	Minimal-risk
Social scoring, manipulative AI, untargeted scraping of facial images, real-time biometric ID in public spaces (with narrow exceptions).	AI in HR, education, essential services, law enforcement, medical devices. Strict obligations: risk management, data quality, logging, human oversight.	Chatbots, deepfakes, emotion recognition. Transparency duties - users must know they are interacting with AI or seeing AI-generated content.	Spam filters, AI in video games, productivity helpers. No specific obligations under the AI Act, but other laws (GDPR, copyright) still apply.

Most everyday uses are minimal-risk - but as soon as you make a chatbot or a deepfake, transparency rules kick in.

GDPR meets AI: the rules still apply

If your AI inputs include identifiable people, GDPR still applies. The six core principles:

Lawfulness

Have a legal basis (consent, contract, legitimate interest)

Purpose Limitation

Use the data only for the original purpose you stated

Data Minimisation

Share only what the AI task actually needs

Accuracy

Do not feed wrong data; correct outputs are wrong

Storage Limitation

Do not keep prompts and outputs longer than needed

Integrity & Confidentiality

Protect against leaks; use enterprise tools when sensitive

Special categories (health, biometrics, voice, religion, politics) need stronger justification - usually explicit consent. **Data subject rights** (access, erasure, objection) apply even to data sitting in an AI tool.

Disclosure and transparency duties

More and more rules require you to say when AI was involved. The default is moving toward “disclose”



When disclosure is usually required:

- AI-generated audio, image, or video resembling real people, places, or events (EU AI Act).
- Chatbots and AI agents interacting with users - make it clear they are not human.
- Academic and journalistic work - most universities, journals, and newsrooms now require an “AI use” note.
- Client deliverables - check your contract or agency policy; “silent” AI use can be a breach.

A simple practice that covers most cases: add a one-line “AI use note” at the bottom of documents - e.g. “Drafted with AI assistance; reviewed and edited by [name].”



● Main activity: Make a request lawful (10 minutes)

You will practise the full skill using fictional examples. Your goal is to turn risky AI uses into lawful ones - without giving up the benefit

For each scenario:

1. Identify the legal risks (personality / copyright / regulation)
2. Rewrite the request lawfully (consent, licensed alternatives, abstraction)
3. Choose the disclosure (AI-use note, label, no need)
4. Decide: proceed / proceed with changes / stop

Output: A short Legal AI Playbook you can use in daily AI work.



Practice 1: Voice clone (personality + disclosure)

Scenario (fictional): “My friend is sick. I want to clone the voice of her favourite singer to record a personalised song for her birthday.”

Write your answers (short):

Legal risks: ____

Lawful rewrite: ____

Disclosure: ____

Decision (proceed / change / stop): ____

Step-by-step tasks:

- **Voice Gate:** Do you have permission from the singer (or licence) for this voice and use?
- **Risk:** Could this be confused with the real singer? Is it commercial use?
- **Lawful rewrite:** Use a generic AI voice or write the song yourself; gift it as clearly AI-made fan content.
- **Disclosure:** Label it clearly as AI-generated when you share it, even with one friend.



Practice 2: Image use (likeness + GDPR)

Scenario (fictional): “I found a photo on Instagram of a stranger's smiling child. I want to use AI to put them on my company's homepage as a 'happy customer'.”

Write your answers (short):

Legal risks: _____

Lawful rewrite: _____

Disclosure: _____

Decision (proceed / change / stop): _____

Step-by-step tasks:

- **Image Checklist:** Right to the photo? Right to the person (and consent for a child)? Context? Disclosure?
- **GDPR:** A child's image is sensitive. You almost always need parental consent and a clear purpose.
- **Lawful rewrite:** Use a licensed stock photo, an AI-generated fictional person clearly labelled, or get explicit consent.
- **Disclosure:** If the image is AI-generated, label it; never present a fictional person as a real customer

Your Legal AI Playbook (keep this for future use)

Write short rules you can apply the next time you use AI.

My voice rule:

"I will only use AI to clone or imitate a real voice if"

My image rule:

"I will not put a real person's face into AI unless"

My disclosure rule:

"When AI helped, I will say"

My copyright rule:

"Before publishing AI outputs, I will check for"

My GDPR / AI Act rule:

"If the input includes other people's data, I will"

“Even if they don’t have copyright ownership over what the AI output is, whoever put that string of queries together and actually was responsible for creating that deepfake, AI or not, still has liability”

Tara Aaron-Stelluto intellectual property lawyer
C.Frawley, What Legal Protections Do Actors Have Against AI?,
<https://www.backstage.com/magazine/article/ai-legal-protections-76366/>



www.valuesai.eu

Follow Our Journey



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

MODULE 08

Final Capstone Module

The AI-assisted apology message



Contents

01	Background	3
02	Instructions	9
03	How to Play	15
04	Capstone Toolkit	19
05	Start Card	21
06	Second-Level Card	24
07	Outcome	31
08	Evidence Capture	50



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

01

Background

AI has moved into our most personal conversations

The Shift

A decade ago, AI helped us spell-check emails. Today it drafts apologies, break-up messages, condolence notes and answers to our parents. The technology is the same - the stakes are very different. When a generator writes on our behalf in an emotional situation, it is not just producing text. It is shaping a relationship.

AI has moved into our most personal conversations

01

Speed

AI drafts in seconds. Repair takes time we no longer give it.

02

Polish

Smooth wording can hide blame, deflection or factual errors

03

Distance

It feels safer to let a machine speak - but the other person still receives the words

Anatomy of a sincere apology

A framework from conflict research (Lazare; Lewicki) - **four moves** that make repair feel real.



01	02	03	04
Acknowledge	Accept Responsibility	Repair	Change
Name what happened, clearly, with the other person's experience in mind.	Own the part you are responsible for. Do not shift blame to the upset person.	Address the impact. Offer something concrete where it is possible and honest.	Describe what will be different. Only promise what you can actually keep.
AVOID <i>"if anyone was upset..."</i>	AVOID <i>"You took it the wrong way"</i>	AVOID <i>Empty gesture or deflection</i>	AVOID <i>"We will never do this again"</i>

Takeaway A short, honest message that hits all four moves usually lands better than a long, polished one that skips any of them.

Where AI helps - and where it fails

Language models are strong at form and weak at truth. That matters more in conflict than almost anywhere else.



Where It Helps

- ✓ **Structure** - Suggests a clear order - acknowledge, impact, next step.
- ✓ **Register** - Calms a hot draft into a calmer tone you can then own.
- ✓ **Options** - Offers alternative phrasings when the right word is stuck.
- ✓ **Accessibility** - Helps non-native speakers and people with dyslexia produce readable text.



Where It Fails

- ✗ **Facts** - Invents details ("hallucination"). In a conflict, invented facts re-open the argument.
- ✗ **Blame-shifting** - Default apology templates often imply it was the other person's reaction, not the act.
- ✗ **Overpromising** - Confident phrasing ("this will never happen again") lands as dishonest when it can't be kept.
- ✗ **Privacy** - Draws in names, screenshots or context that should not be published in a public chat.

Four guardrails for AI-assisted messages

Before you send anything, a machine helped you write, check it against these four.

Verify



V

Every factual claim about what happened must be checked with the people involved - not with the AI.

Ask yourself: *Would I stake my reputation on this sentence being accurate?*

Fairness



F

Consider who is visible, who is missing and who could be harmed if this message is screenshot and forwarded.

Ask yourself: *Would the person I'm writing about feel this is a fair summary?*

Privacy

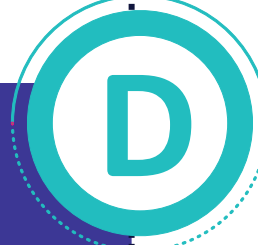


P

Names, identifying details and third-party information should be removed unless there is a clear reason to keep them.

Ask yourself: *Does this message share something that is not mine to share?*

Disclosure



D

In a conflict, being honest about AI assistance builds trust; hiding it often breaks trust if it later comes out.

Ask yourself: *Would it feel wrong if the recipient learned AI wrote part of this?*

02

Instructions

The AI-Assisted Apology Message

An interactive decision game. Read a scenario, apply your tools, make a choice and see what happens.



Scenario



A group-chat apology



Paths



A · B · C



Outcomes



Discussion + reflection

The scenario & the golden rule

**Set the stage
before you make
a single click.**



The Scenario -A disagreement in your youth group has spilled into a group chat.

A message was posted that upset someone. A leader asks your group to help draft an apology that will be shared in the chat.

Your task: decide how to handle the apology: with AI, without AI, or somewhere in between.

The scenario & the golden rule

**Set the stage
before you make
a single click.**



Your Toolkit - pick any two

- STOP - THINK - CHECK - CHOOSE
- THINK: True · Helpful · Intent · Neutral · Kind
- Fairness check - who is included, missing, harmed?
- Verification habits - pause, check details, check source
- Pause · Verify · Protect · Support · Report

Golden Rule

Use at least two tools before every decision.

How a round works

- five steps

Read → apply tools → choose → read outcome → reflect.



01

Read the start card

Look at the scenario on card B0. Notice the decision prompts on the right.

02

Pick a path: A, B or C

A = use AI, send it.
B = AI draft + edit.
C = write it yourselves.

03

Apply two tools

Before clicking, name the two toolkit items you are using and why.

04

Read the outcome

See what happens in the chat and discuss the three open questions.

05

Capture reflection

On the reflection slide, record your path, your tools and a safer follow-up action.

TIP Different paths lead to very different outcomes – compare choices with your group before moving on to the reflection slide.

Module 8 Capstone

AI Decision Game

The AI-assisted apology message

[Start](#)



03

How to play

Navigate the scenario like a digital decision game

This deck is set up for classroom discussion, self-study or live facilitation.



Flow

1. Read the start card
2. Pick a route: A, B or C
3. Apply at least two tools
4. Click into the next card
5. Read the outcome and discuss the open questions
6. Capture evidence on the reflection slide

Navigate the scenario like a digital decision game

This deck is set up for classroom discussion, self-study or live facilitation.



Suggested facilitation moves

- Ask learners to name the two tools they used before clicking
- Pause after each outcome and compare different paths
- Invite learners to edit the wording live and explain why
- Use the reflection slide as an evidence record or exit task

Navigate the scenario like a digital decision game

This deck is set up for classroom discussion, self-study or live facilitation.



Jump into the toolkit first if you want the prompts visible before starting.

Deck Map

M B0 Start card



B-A, B-B or B-C



Outcome cards



Reflection

[Open toolkit](#)

04

Capstone Toolkit

Decision tools to apply before each choice

Use at least two prompts every time you make a decision in the game.



Core prompts

- STOP–THINK–CHECK–CHOOSE
- THINK = True, Helpful, Intent, Neutral, Kind
- Fairness check = who is included, who is missing, who could be harmed
- Verification habits = pause, check details, check source

Extended prompts

- Pause–Verify–Protect–Support–Report
- Create → Verify → Fairness check → Disclose
- NSL, Consent Gate, Fair Data Check, SAFE Workflow Ladder

05

Start Card

The Apology Draft



Scenario

A disagreement in your youth group has spilled into a group chat. A message was posted that upset someone. A leader asks your group to help draft an apology message that will be shared in the chat.

The Apology Draft



Choose a path. Each button links to the next scenario or outcome card.

Decision prompts

- Pause before posting
- Check what is accurate
- Consider privacy and fairness

Choice A

Use AI to write the apology and send it straight away.

Choice B

Use AI to draft, then edit it in your own words and check the details first.

Choice C

Do not use AI. Write it yourselves to avoid mistakes.

06

Second-Level Card

The AI message looks polished



Scenario

The AI draft sounds calm and confident. It includes one line about what happened that you are not sure is accurate. It also names the person who was upset.

The AI message looks polished

Choose a path. Each button links to the next scenario or outcome card.

Decision prompts

- Verification first
- Do not name people casually
- Shared facts matter

Choice A1

Send it as written.

Choice A2

Remove the name and the disputed line, then send it.

Choice A3

Pause and ask the group to confirm what actually happened before sending anything.

Editing the Draft



Scenario

You use AI to draft, then you start editing. The AI draft includes a phrase that sounds like it is blaming the upset person (“if you took it the wrong way...”). It also includes a promise (“we will never do this again”) that you cannot guarantee.

Editing the Draft

Choose a path. Each button links to the next scenario or outcome card.

Decision prompts

- Check tone for blame
- Do not promise what you cannot keep
- Choose a calmer process

Choice B1

Keep the phrase and promise because it sounds professional.

Choice B2

Rewrite the message to focus on impact and remove promises you cannot keep.

Choice B3

Share the draft with the whole chat first and ask people to vote on it.

Writing without AI



Scenario

You write your own apology. It is short and honest, but you are worried it might sound too blunt. Someone suggests adding an AI-generated line to “smooth it out”.

Writing without AI

Choose a path. Each button links to the next scenario or outcome card.

Decision prompts

- Clarity matters
- Tone still needs checking
- AI should not replace accuracy

Choice C1

Keep it short and send it as-is.

Choice C2

Add a single AI-suggested sentence, then edit it to match your voice.

Choice C3

Ask AI to rewrite the whole message to sound more persuasive.

07

Outcome

Sent quickly, challenged publicly

Outcome card and debrief prompts



What happens

The message is posted quickly. Some people respond positively. Others reply asking about the disputed line. The named person reacts in the chat and the conversation becomes active again.

Discuss, annotate or jump to the reflection slide to capture evidence.

See Open Questions on next slide



Open Questions

Sent quickly, challenged publicly

1. Which details would you want to verify before sending a message like this?
2. What is the risk of naming someone in a public group chat?
3. What would you do if someone replies with a different version of events?

Edited for privacy, but still prompts questions

Outcome card and debrief prompts



What happens

The message is posted without naming anyone. Some people accept it. A few ask for more clarity about what happened. The person who was upset reads the message but does not respond immediately.

Discuss, annotate or jump to the reflection slide to capture evidence.

See Open Questions on next slide



Open Questions

Edited for privacy, but still prompts questions

1. What information is necessary in an apology and what is optional?
2. How can you acknowledge impact without restarting the argument?
3. What would you do next if the upset person messages privately?

Shared facts before posting

Outcome card and debrief prompts



What happens

You do not post yet. You ask for a quick check of what happened. People share different recollections. The group agrees on a short, shared summary before drafting the apology.

Discuss, annotate or jump to the reflection slide to capture evidence.

See Open Questions on next slide



Open Questions

Shared facts before posting

1. What is a fair way to agree a shared summary when people disagree?
2. How can you keep the process calm and respectful?
3. When is it better to move the conversation out of the group chat?

Professional tone, unsafe wording

Outcome card and debrief prompts



What happens

The edited message is posted with the blaming phrase and a strong promise. Some people react with thumbs-up. One person replies pointing out the blaming wording. Another says the promise sounds unrealistic.

Discuss, annotate or jump to the reflection slide to capture evidence.

See Open Questions on next slide



Open Questions

Professional tone, unsafe wording

1. Which part of the message could be read as blame and why?
2. What promises are safe to make and which are risky?
3. How could you rewrite the message to focus on impact and repair?

Impact-focused repair

Outcome card and debrief prompts



What happens

The final message focuses on impact and avoids blame. It includes one clear next step, for example “we will keep the chat respectful and check before we share”. Several people respond positively and the chat slows down.

Discuss, annotate or jump to the reflection slide to capture evidence.

See Open Questions on next slide



Open Questions

Impact-focused repair

1. What makes a repair message feel sincere rather than scripted?
2. Which tools helped you decide what to change?
3. What is one follow-up action that supports the words in the apology?

Crowdsourcing brings the argument back

Outcome card and
debrief prompts



What happens

You share the draft to the whole chat for a vote. People comment on wording, tone and details. Some suggestions conflict. The conversation becomes long and the original issue reappears in replies.

Discuss, annotate or jump to the reflection slide to capture evidence.

See Open Questions on next slide



Open Questions

Crowdsourcing brings the argument back

1. What are the risks of crowdsourcing wording in a live group chat?
2. What alternative process could reduce conflict and still include voices?
3. What would you do if someone shares screenshots outside the group?

Honest but too brief for some readers

Outcome card and debrief prompts



What happens

You send the short message. Some people appreciate it. A few people interpret it as cold or too brief. The person who was upset replies asking for a clearer acknowledgement of impact.

Discuss, annotate or jump to the reflection slide to capture evidence.

See Open Questions on next slide



Open Questions

Honest but too brief for some readers

1. What information or wording might make this feel more caring?
2. How can you be clear and kind at the same time?
3. What would you add without turning it into a long explanation?

Light AI support, voice retained

Outcome card and debrief prompts



What happens

You add one AI-suggested sentence, then edit it. The final message stays short but includes a clearer acknowledgement of impact. The person who was upset responds with a short reply and the chat quietyens.

Discuss, annotate or jump to the reflection slide to capture evidence.

See Open Questions on next slide



Open Questions

Light AI support, voice retained

1. What is one sign that an AI sentence needs editing before sending?
2. What would you disclose, if anything, about AI assistance in a group message?
3. How would you check that the tone is fair to everyone involved?

Persuasive rewrite, shaky facts

Outcome card and debrief prompts



What happens

You ask AI to rewrite the whole message. It becomes longer and more persuasive. It includes a phrase that is not accurate to what happened. Some people challenge the message in replies.

Discuss, annotate or jump to the reflection slide to capture evidence.

See Open Questions on next slide



Open Questions

Persuasive rewrite, shaky facts

1. What should you verify when AI rewrites a personal situation?
2. How can persuasive wording create problems in a conflict situation?
3. What would you do next if people accuse the message of being “fake” or “scripted”?

08

Evidence Capture

Quick Reflection & Portfolio Prompts

Use this slide as a learner record, plenary prompt or screenshot template.

Record your reasoning

- Which path did you choose?
- Which two tools did you use?
- What did you verify?
- Who could be affected or excluded?
- What would you disclose about AI use, if anything?
- What is one safer follow-up action?

Quick Reflection & Portfolio Prompts

Use this slide as a learner record, plenary prompt or screenshot template.

Facilitator Prompts

- What made the message feel sincere or unsafe?
- Where did accuracy matter most?
- What changed after the fairness check?
- When should the conversation move out of the public group chat?
- What evidence would show responsible AI use here?



www.valuesai.eu

Follow Our Journey



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.