EXChange AI: Module 3 · AI and Society

TEACHING NOTES FOR EDUCATORS

PROJECT 2024-2-BG01-KA210-ADU-000296153

These teaching notes accompany the **Expeditions for Change: Navigating the Skills & AI Landscape** project. Each section corresponds to a slide and functions as a teaching script — what to explain, how to transition, and how to activate your audience. Visit **exchangeai.eu** for full project resources.

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Human Resource Development Centre. Neither the European Union nor the granting authority can be held responsible for them.

Slide 1 & 2 — Setting the Stage: The Goldilocks Rule

Opening the Module

Introduce the theme: not only *what* AI is, but how it already affects society, work, education, trust, and decision-making. Emphasise a **realistic, balanced view** — neither advertising for AI nor a fear-based picture.

Connect to adult learners: AI literacy means understanding social consequences — bias, misuse, safety, jobs, and environmental footprint. Invite critical thinking throughout.

The Goldilocks Metaphor

Public discussion swings between two extremes:

- **Excessive optimism:** AI will solve every problem
- **Excessive pessimism:** AI will inevitably lead to catastrophe

The **third position**: AI is a powerful tool — not magic, not an independent moral agent. It is created, trained, deployed, and controlled by people and organisations. The focus is **responsible use**: increasing value and reducing harm.

Ask participants: "Which of these three positions do you encounter most often in the media or among colleagues?"

Slide 3 — AI Limitations: Performance & Explainability

Every AI system has limitations, even when results look impressive. Use the medical imaging example — systems trained to recognise patterns like pneumothorax in X-ray images — to illustrate both the strength and the fragility of AI.

Performance Limitations

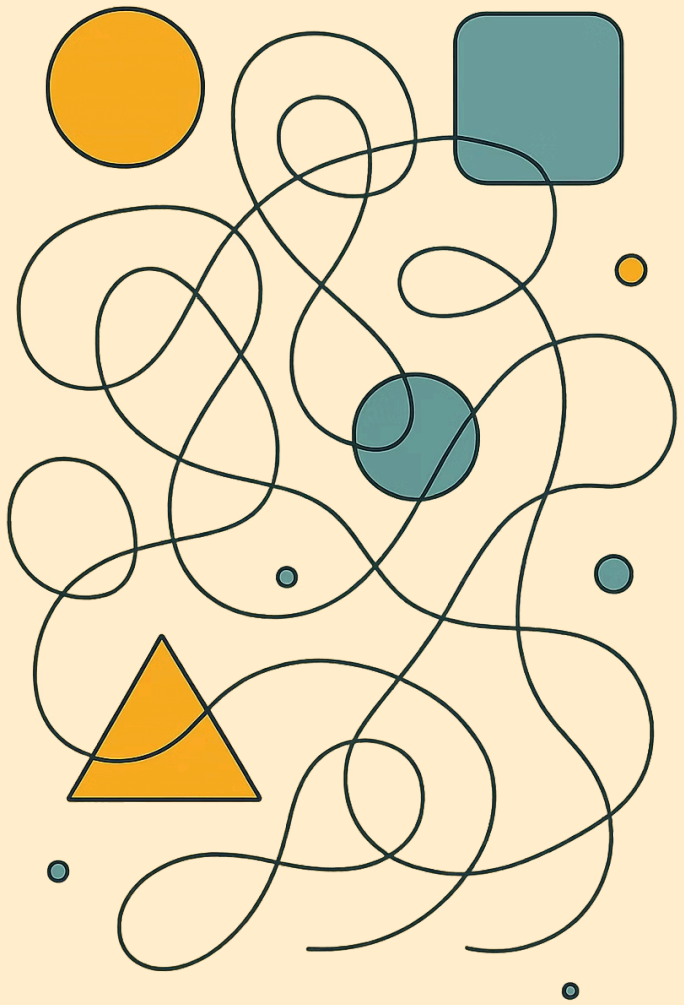
A system may perform well in test conditions but make mistakes with different data, different patients, or a different context. High accuracy on benchmarks does not guarantee real-world reliability.

Explainability Limitations

It is sometimes difficult to understand *why* a model produced a specific result. In sensitive areas — medicine, finance, education — it is not enough to have an answer; we also need to check its logic and reliability.

The Core Principle

AI should **support**, not uncritically replace, professional judgement. Human oversight remains essential wherever decisions carry real consequences.



Part 1 — Discrimination & Bias in AI

AI does not emerge in a vacuum. It is trained on data created by people and societies — and these data often contain **historical inequalities, stereotypes, and incomplete representation** of different groups. Prepare the audience to view bias not as a "computer error," but as a social problem that can be amplified through automation.

Slides 4–7

This section covers how bias enters AI systems, its real-world consequences, and what can be done — technically, organisationally, and socially.

Why It Matters for Educators

Education should promote fairness, inclusion, and critical thinking. Understanding bias is essential for any educator using or recommending AI tools.

Slides 5–6 — How Bias Works & Why It Matters

How AI Learns Stereotypes (Slide 5)

Use the word embedding analogy: a system may learn links such as **"man — computer programmer"** and **"woman — homemaker"** — not because this is true or desirable, but because such associations were common in training texts.

The problem is not only linguistic. When such associations are embedded in real systems — search engines, chatbots, recruitment tools, or educational platforms — they can **influence decisions and recommendations**.

> ⚠ The model does not "understand" fairness — it recognises statistical patterns. Human review, data design, and ethical oversight are therefore essential.

Invite participants to give examples of stereotypes that often appear in professional or educational language.

Real Consequences at Scale (Slide 6)

Bias matters because AI systems are increasingly used in **high-stakes processes**:

- Candidate selection & recruitment
- Facial recognition
- Lending & credit scoring
- Risk assessment
- Service personalisation

The harm is not only individual but **systemic**. At scale, entire groups can be placed at a disadvantage. Encourage participants to ask:

- Who is represented in the data?
- Who is missing?
- Who checks the outputs?
- What are the consequences of an error?

Slide 7 — Combating Bias: A Multi-Layer Approach

Combating bias requires a combination of **technical, organisational, and social measures**. Technology alone is not enough.

Technical Measures

- Reducing or "zeroing out" biases in word embeddings
- More careful data selection and curation
- Testing across different demographic groups
- Periodic audits of model outputs

Organisational Measures

- Transparency in how systems are built and deployed
- Clear accountability procedures
- Diverse development teams who can spot invisible problems
- Ethical review boards and oversight processes

Educator Practice

- Check AI outputs before sharing with learners
- Compare perspectives from multiple sources
- Avoid presenting automated answers as neutral truth
- Model critical evaluation for participants

Part 2 — Adversarial Attacks on AI (Slides 8–12)

Here we move from unintended bias to **deliberate manipulation**. Adversarial attacks involve small, intentional changes to data or the environment that can mislead a model — changes that may be nearly invisible to a human but cause the system to produce a completely different classification.

1

Digital Attacks

Tiny pixel-level perturbations added to images cause dramatic misclassification — e.g., "hummingbird" becomes "hammer."

2

Physical Attacks

Glasses, stickers, patches, or altered signs confuse recognition systems in the real world — stop signs, faces, access control.

3

Confident Wrong Answers

Systems often don't say "I don't know" — they give a confident but incorrect answer. Confidence ≠ truth.

This topic is important because many AI systems operate in the real world: cameras, autonomous vehicles, biometric identification, content filters, and security systems. The more critical the application, the more important it is to understand how it can be attacked.

Slide 12 — Defences Against Adversarial Attacks

Available Defences

- **Adversarial training:** exposing models to attack examples during training
- **Stricter testing:** evaluating robustness beyond standard benchmarks
- **Input filtering:** detecting and blocking suspicious inputs
- **Architectural improvements:** building more robust model structures
- **Continuous monitoring:** watching for anomalous outputs in deployment

The Arms Race Analogy

Every defence has a cost — more computing resources, slower deployment, lower flexibility, or new trade-offs in accuracy.

Like spam and anti-spam: when defences improve, attackers also adapt. In some areas there may be a **continuous arms race**.

Key message: AI security is not a one-time setting, but a process of continuous testing, monitoring, and updating.

Ask: "Where would we be most concerned about such a vulnerability — in social media, cars, banks, or schools?"

Part 3 — Adverse Uses of AI (Slides 13–14)

Make a clear distinction: the previous sections discussed systems that *unintentionally* inherit bias or can be fooled. Here the topic is how people can **deliberately use AI for fraud, manipulation, surveillance, or disinformation**. AI literacy includes not only the ability to create content, but also the ability to recognise, verify, and critically evaluate it.

Deepfakes

Video or audio in which a person appears to say or do something they never did — undermining trust in media, institutions, and personal communication.

Fake Public Opinion

Mass generation of fake comments can create the illusion of widespread public sentiment, distorting democratic discourse.

Automated Scams

AI-powered scams become more convincing and personalised, making them harder to detect and easier to scale.

AI-Enabled Surveillance

Facial recognition and behaviour analysis can restrict freedoms and rights in oppressive contexts, even when the technology itself is neutral.

Encourage educators to include **media literacy practices**: checking sources, searching for context, recognising manipulative signals, and sharing content carefully.

Part 4 — AI, Jobs & the Planet (Slides 15–20)

AI is not only a technological issue — it has profound **economic and environmental dimensions**. Adult learners and working professionals are precisely the groups that will need reskilling, additional competences, and support for adaptation. This is why the role of educators is crucial.

The Jobs Dimension (Slides 15–17)

By 2030, automation may affect hundreds of millions of jobs — while new jobs are also expected to be created. The key framing is **transformation, not just replacement**: many professions will not disappear entirely, but the tasks within them will change.

Generative AI increases automation potential because it can now work with text, images, code, and communication tasks previously considered harder to automate. Occupational groups most affected include office support, production work, services, customer service, and business and legal activities.

Ask: "Which tasks in your work can be supported by AI, and which should remain under human control?"

Human Skills That Remain Essential

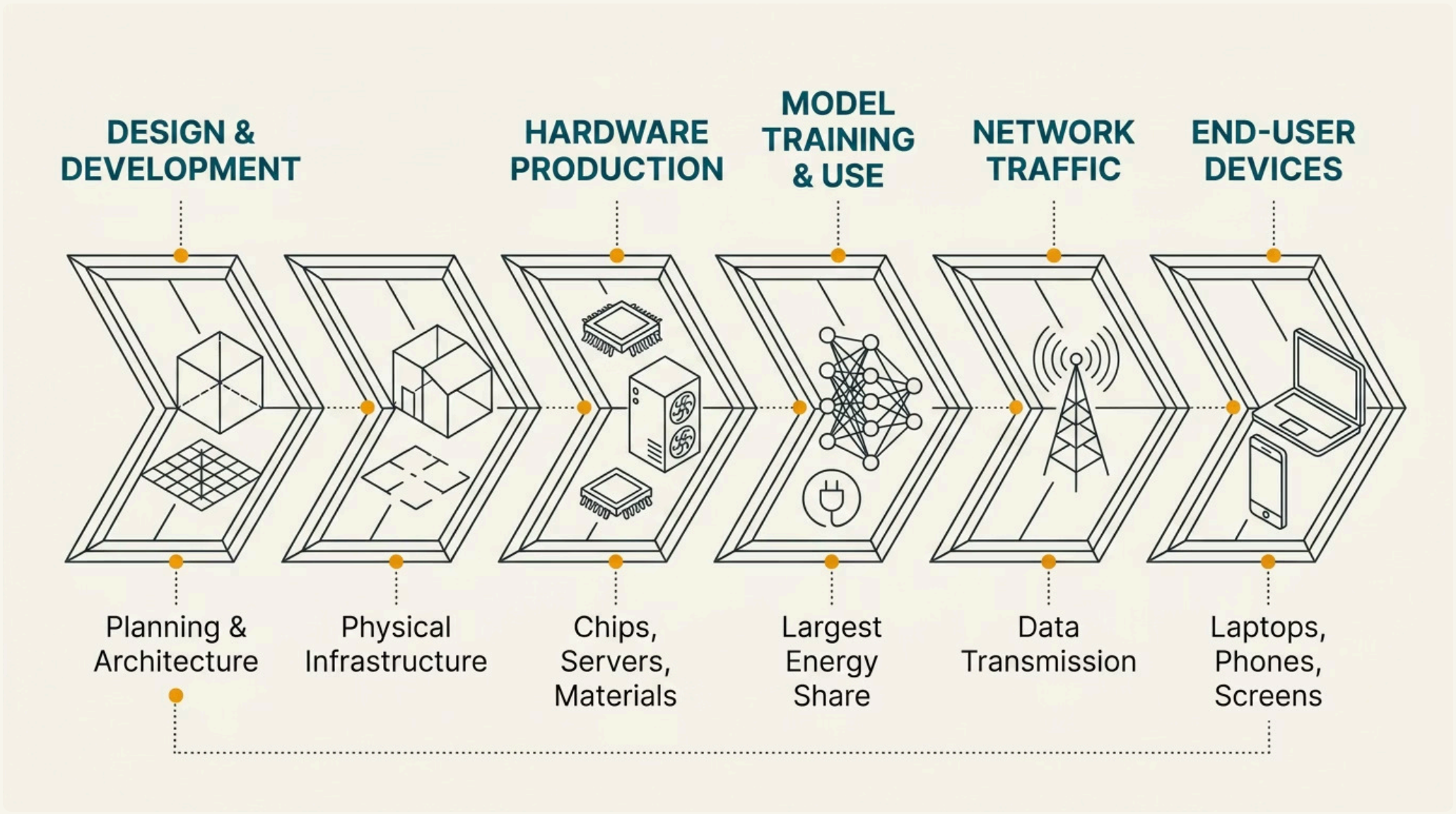
- Critical thinking and ethical reasoning
- Communication and interpersonal skills
- Pedagogical design and facilitation
- Working with and for people
- Output verification and quality judgement

Policy Responses (Slide 18)

Solutions combine individual, organisational, and policy measures. Conditional basic income is one example of a safety net linked to learning incentives. **Lifelong learning** is central: people cannot rely only on knowledge from initial education. Organisations and governments must support continuous skills updating.

Slides 19–20 — The Environmental Footprint of AI

Large AI models have a full life cycle — and each stage carries a real footprint in terms of emissions, water, and materials. **Digital services are not "weightless."**



What the Data Shows (Slide 19)

A large share of impact comes from model training and use, as well as from hardware components. The location of data centres, the energy mix, and the size of the model all matter significantly.

The key message: this does not mean we should stop using AI, but that we should use it **consciously and proportionately to the task**.

Making It Concrete (Slide 20)

Even generating one page of text has a small but real footprint — emissions, water, and materials. Teach participants **efficient AI use**:

- Write clear, precise prompts
- Group questions rather than repeating
- Avoid unnecessary generation of multiple versions
- Choose the appropriate tool for the task
- Don't use large models for trivial tasks when a lighter solution exists

This combines **productivity with responsibility**.

Slide 21 — Module Conclusion: AI Literacy as Civic Competence

Conclude by summarising the central idea: **AI literacy is not only a technical skill, but also a civic, professional, and ethical competence.** Participants should be able to use AI — but also understand its limitations, risks, possible misuse, and social consequences.

1

Bias & Discrimination

AI inherits and can amplify historical inequalities. Human oversight, diverse teams, and ethical design are essential.

2

Adversarial Attacks & Security

Deliberate small changes can fool AI systems. Security is a continuous process, not a one-time setting.

3

Misuse & Disinformation

Deepfakes, fake opinion, and surveillance require media literacy and critical evaluation skills alongside AI use.

4

Work & the Planet

AI transforms jobs and carries environmental costs. Lifelong learning and conscious use are the educator's response.

Encourage educators to introduce AI literacy into their courses through discussions, practical tasks, case analysis, and critical reflection. Visit **exchangeai.eu** for the remaining modules and all project resources.

Visit exchangeai.eu