DW Akademie

MEDIA AND INFORMATION LITERACY

Heroes and Villains in the age of AI

Digging deeper into AI concepts and technical background

ANNEX A

Annex A: digging deeper into AI concepts and technical background

Purpose
The core text explains AI in clear, minimal terms so facilitators can guide sessions without technical overload. This annex is the deeper layer behind those explanations. It expands the "why" and "how" of AI behavior so you can answer questions more conf< # >idently, spot misunderstandings quickly, and frame discussions with balance.

You are not expected to teach this annex.
It is reference material: use it when you want stronger grounding, more nuance, or better language for difficult questions.

1. Artificial intelligence is not new

AI is not new and not all AI are the same- depending on their stage of development, they differ in terms of what they can do and how they work.
The term Artificial Intelligence was coined in 1956 and the concept has evolved greatly since then. Mostly the concept of AI refers to technical systems which do tasks which normally require human intelligence. General public use of AI exploded in 2023 with the release of Chat GPT which had an interactive user interface and generated new content.
Let's take a quick look at some key terms:

AI Models- are computer programmes that learn from examples. They recognize patterns and connections in data to create predictions.

For example: *It can learn "if there are dark clouds, it is likely to rain". The model does not understand the causes- it calculates probability. Because it learns from the available data, it can take in errors or biases and supply results which are incorrect under new conditions.*

Examples of AI Models:
I*mage Recognition Models (recognizing and analyzing images like face recognition on your phone)*
Binary Classification Models (deciding between two options like if a movie was reviewed positive or negative)
Generative Models (creating new content like text, images, videos or audio)

AI Systems- are a complete solution that integrate one or more AI models into a functional framework to solve real world problems and provides the user interface.

For example: *A self driving car is an AI system that uses several different AI models (one for vision, one for steering, one for signs) to achieve its goal.*

Generative AI- can create something new- for example, texts, videos, images, speech or music.
For example: *Using CHAT GPT or Claude to write an email*

Where AI shows up in everyday life

AI appears in different ways across daily media environments.

- **Embedded AI** works in the background of social media feeds, messaging apps, search engines, and streaming platforms.
- **Direct AI tools** are used intentionally, such as chatbots, translation tools, or image generators.
- **Hidden AI systems** shape recommendations, rankings, filtering, and moderation without always being visible to users.

The level of control varies depending on how AI is encountered. Some systems are chosen directly. Others are built into platforms people use every day.

A possible consequence is that people may not always realise when AI is influencing what they see, search for, or engage with.

Example: A chatbot allows users to choose prompts and settings, while a social media feed is often shaped automatically.

Example: A user may believe they are seeing the "most important" news stories, without realising ranking systems shaped what appeared first.

2. Pattern learning vs. understanding

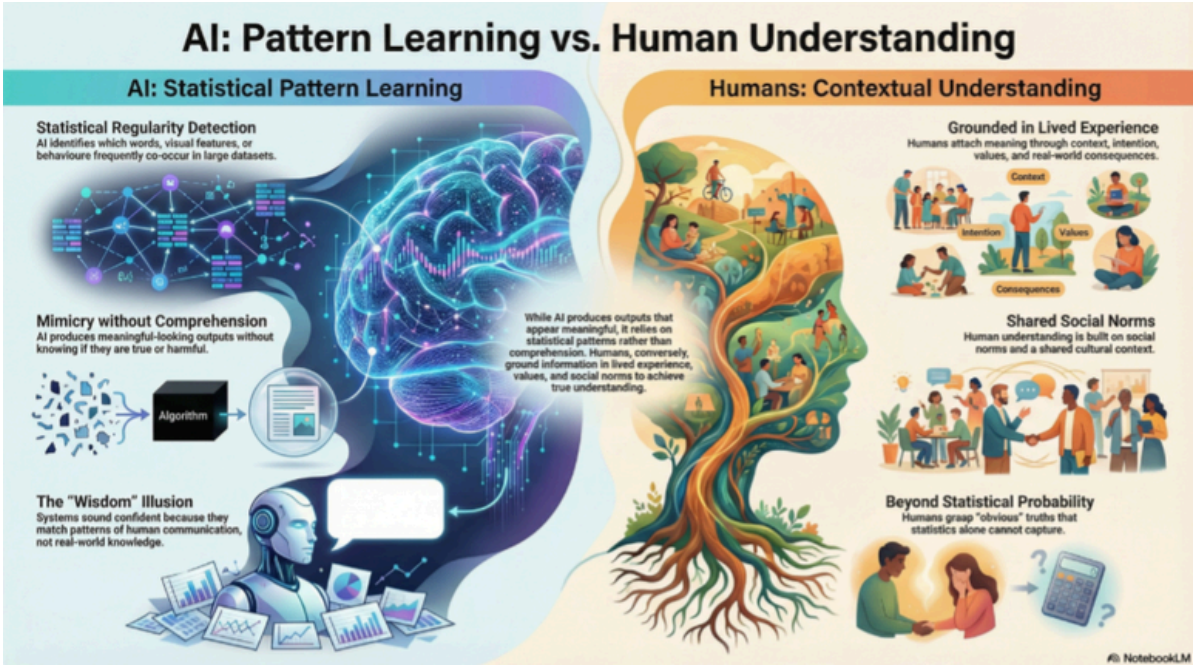
A key idea in this supplement is that AI does not "understand" in a human sense. Humans attach meaning to information through context, intention, values, and lived experience. We connect words and images to real-world knowledge, consequences, and social norms. AI does none of this.

AI systems learn by exposure to huge numbers of examples. They detect statistical regularities: which words tend to appear near each other, which visual features often co-occur, which user behaviors often predict another click. This is pattern learning, not comprehension.

The practical consequence is important: AI can produce outputs that look meaningful because they match patterns of meaningful human communication. But it does not know what its output refers to, whether it is true, whether it could harm someone, or how it will be interpreted in a specific cultural moment.

This is why AI can be impressive and misleading at the same time. It can sound wise and confident without being grounded. It may also miss what humans find obvious—because "obvious" often depends on real-world knowledge and shared context, not on statistical similarity alone.

***Example**: A chatbot can write a warm apology message that sounds sincere, even though it has no idea what happened, who was hurt, or what the real consequences are.*



*AI Generated with Google Notebook LLM

3. AI as a prediction system

Modern AI is best understood as a prediction system. Across different applications, it learns to produce the most likely output given an input—based on patterns observed in training data.

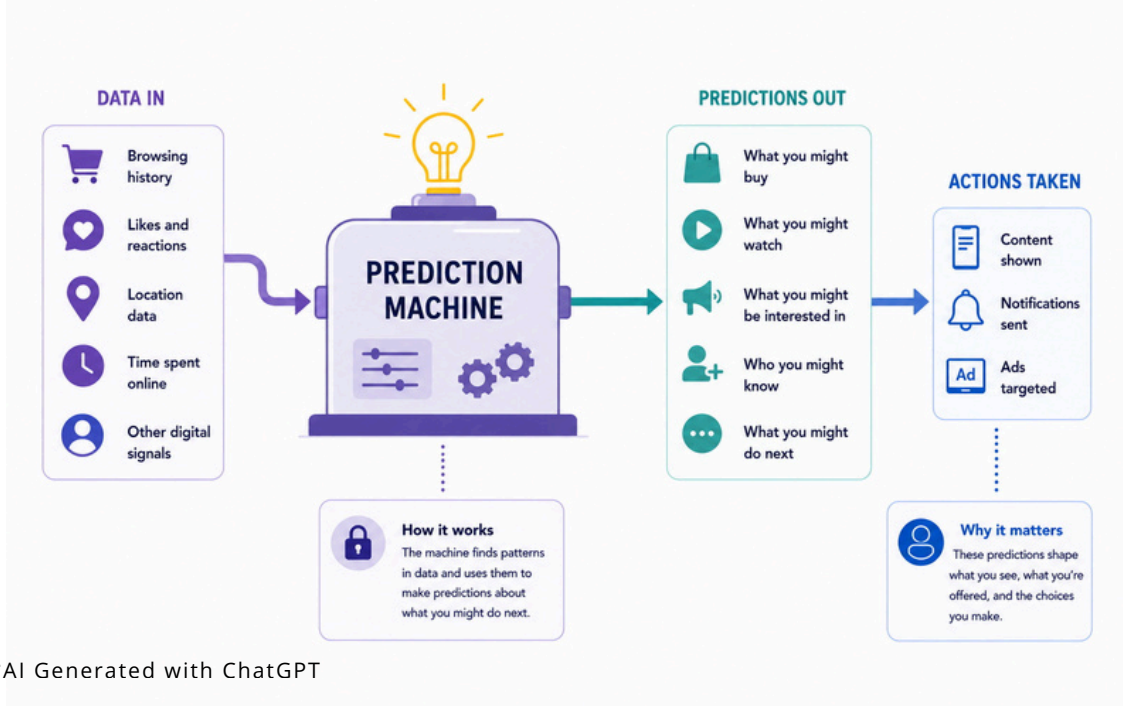
This prediction principle shows up in different forms:

- Text models predict the next word (or the next chunk of text).
- Image generators predict what visual elements should appear to match a prompt.
- Recommender systems predict what you might click, watch, or linger on next.
- Classification systems predict which label best fits content (spam/not spam, safe/not safe, category A/B/C).

Prediction can look like intelligence because it often produces coherent results. But prediction is not the same as reasoning. The system does not "decide" in a reflective sense. It does not check whether a result is responsible, appropriate, or fair—unless humans have added specific constraints and feedback loops to encourage that behavior.

In facilitation terms: this is why "AI feels smart" is not a silly reaction. It's a human interpretation of fluent prediction. Your job is not to shame that reaction, but to reframe it: "It sounds smart because it predicts well—not because it understands."

Example: Autocomplete in your phone predicts the next word so well that it feels like it "knows" what you mean—yet it's only guessing based on patterns from past text.



*AI Generated with ChatGPT

4. Why AI hallucinates

"Hallucination" describes a predictable failure mode: AI generates information that sounds plausible but is inaccurate or invented. This is not rare, and it is not always obvious. The system may produce correct and incorrect statements in the same paragraph with the same confident tone.

Why it happens: if the model is designed to generate the most likely continuation, it will generate something even when it does not truly "know." If training data contains partial patterns or conflicting information, the model may stitch them together into a new, smooth-sounding answer. If the prompt is ambiguous, it may choose one likely interpretation and build on it confidently.

Hallucinations are also more likely when:

- the question asks for precise facts (dates, names, citations) without clear grounding,
- the request pushes beyond the model's training coverage,
- the system is asked to fill in missing details ("make it more specific"),
- the topic is niche, fast-changing, or poorly represented in training data.

To the AI, hallucination is just prediction. There is no internal alarm that signals "uncertain" unless a system is engineered to express uncertainty. This is why facilitator framing matters. Participants should learn a practical habit: treat AI output as draft material until verified - especially when the stakes are high.

Example: *If you ask an AI for the "exact date" of a small local event, it may confidently invent a date that sounds right - because it's filling a gap, not checking a calendar.*

5. Training data - Where AI behavior comes from

AI behavior is shaped by training data the way a person's worldview is shaped by what they have been exposed to - except AI has no values or reflection to correct for what it absorbs. Training data can include books, websites, news, social media, images, videos, and many other sources. The model learns regularities from these sources: not only language patterns, but also social patterns - how groups are described, which voices are present, what perspectives dominate, what is treated as "normal."

Training data quality and balance matter. If certain communities, languages, or cultural contexts are underrepresented, the model may perform poorly for them or misrepresent them. If stereotypes appear frequently, the model may reproduce them. If misinformation is included, the model may echo misleading narratives.

A key global dynamic is that much of the training material used by major AI systems is:

- in majority languages (especially English),
- originating from the Global North or global platforms where those voices dominate,
- formatted and accessible in ways that are easier to scrape and process,
- influenced by the location and incentives of large technology companies.

This does not mean AI can't work in other contexts - but it does mean "what AI knows" and "how AI speaks" can reflect unequal data visibility. For MIL work, this matters because participants may experience AI as culturally off, linguistically awkward, or biased in ways tied to representation - not just to "bad technology."

Example: *A translation tool may handle English extremely well but struggle with a smaller language or dialect, because it has seen far fewer high-quality examples in that form.*

6. Why AI can be biased

Bias in AI is usually not intentional. It is structural: AI learns from data patterns, and data patterns reflect power, history, inequality, and omission. AI cannot evaluate fairness unless humans explicitly build mechanisms to address it - and even then it is difficult.

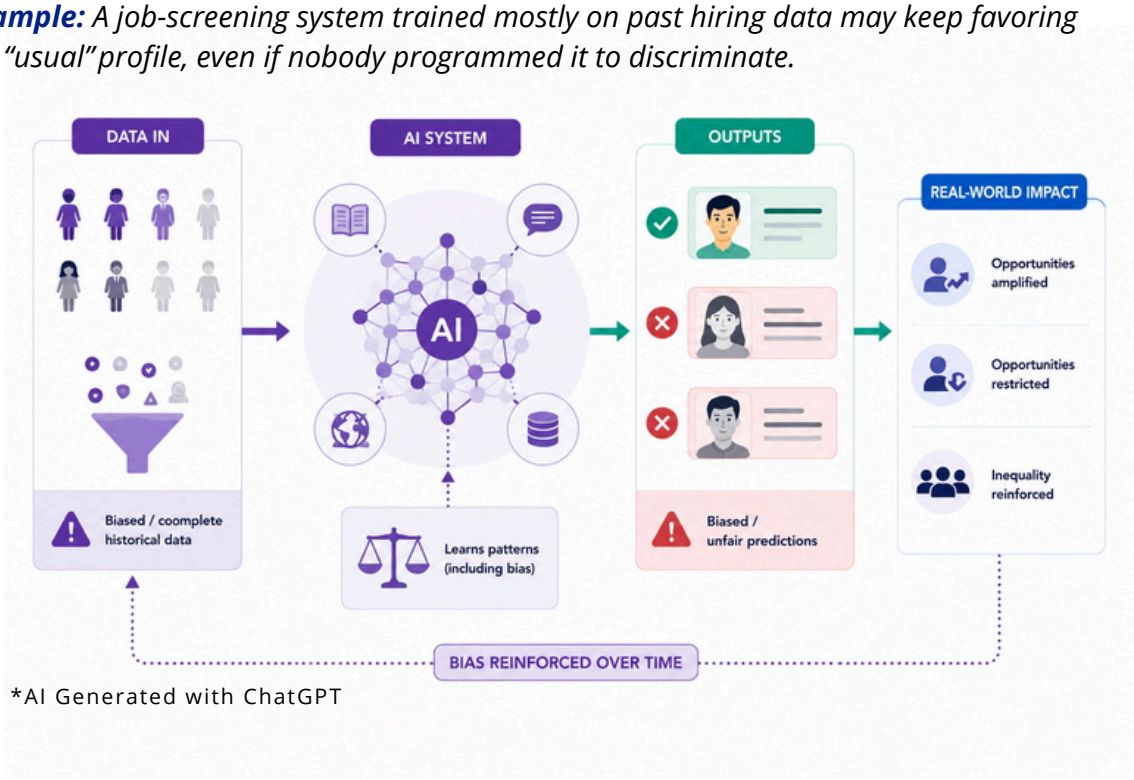
Bias emerges when:

- harmful patterns repeat (stereotypes, exclusions, hostile framing),
- groups are underrepresented (less data → weaker performance),
- historical inequalities are embedded (past injustice becomes "training reality"),
- the optimization goal rewards the wrong thing (engagement, speed, cost reduction).

A critical point for facilitation: bias is not only about offensive outputs. It can also show up as invisible unfairness - who gets flagged by moderation, who is recognized by face detection, whose language is misinterpreted, who is excluded by "neutral" systems.

Bias is not always intentional. AI often reproduces unfairness because it learns from unequal data, historical patterns, and flawed systems. But bias can also be reinforced deliberately when governments, political actors, or platforms use AI to monitor, censor, manipulate, or disadvantage certain groups.

At scale, these harms become powerful in everyday life. This links directly to MIL: participants should ask not only "What did it output?" but also "Whose reality is missing, who benefits, and who may be harmed?"
Example: *A job-screening system trained mostly on past hiring data may keep favoring the "usual" profile, even if nobody programmed it to discriminate.*



*AI Generated with ChatGPT

7. Large Language Models (LLMs)

Large Language Models generate text by predicting what usually comes next in language. They learn from massive datasets of text and absorb patterns of explanation, tone, style, and framing.
This gives them strengths that feel human-like: they can write fluently, summarize, draft, translate, generate examples, and support brainstorming.

But their limitations are equally important. They can produce false claims confidently, mix accurate and inaccurate information, mirror bias in training data, appear empathetic without true understanding, and cannot reliably cite sources unless connected to external retrieval tools.

LLMs are best treated as powerful writing and thinking assistants—not truth machines. Use them for drafts, ideas, and language support, verify important facts externally, and avoid treating them as final authority on sensitive financial, legal, medical, or emotional topics.
This helps participants avoid two extremes: seeing AI as magic or seeing it as useless. It is useful—but not trustworthy in the way people often assume.
Example: An AI can draft a clear email in seconds, but you still need to check names, dates, and factual details before sending it.

AI as a conversational technology

For many people, AI no longer feels like software—it feels like a conversation.
Chatbots can sound personal, supportive, persuasive, and emotionally aware. This can increase trust and sometimes create emotional attachment. Some people already use AI tools for companionship, advice, emotional support, or even as substitutes for human interaction.
AI may feel human in conversation—but it does not think, care, or take responsibility.
The risk is that people may rely on AI for emotional guidance, relationships, or sensitive decisions in ways that create new vulnerabilities.
Example: A chatbot may offer relationship advice or emotional reassurance without understanding real-world consequences.
Example: A young person may treat an AI companion as a trusted friend while receiving harmful or unhealthy advice.

Prompting tip!
It does help to ensure that the prompts you are using are clear and include:

- A role
- A task
- Context
- And an example (if possible)

Example prompt: *Draft an invitation to my friends for a fun get together to celebrate my birthday on March 20th at 18:00 at the community center. The invitation should invite everyone in a friendly, upbeat way.*

8. Generative AI

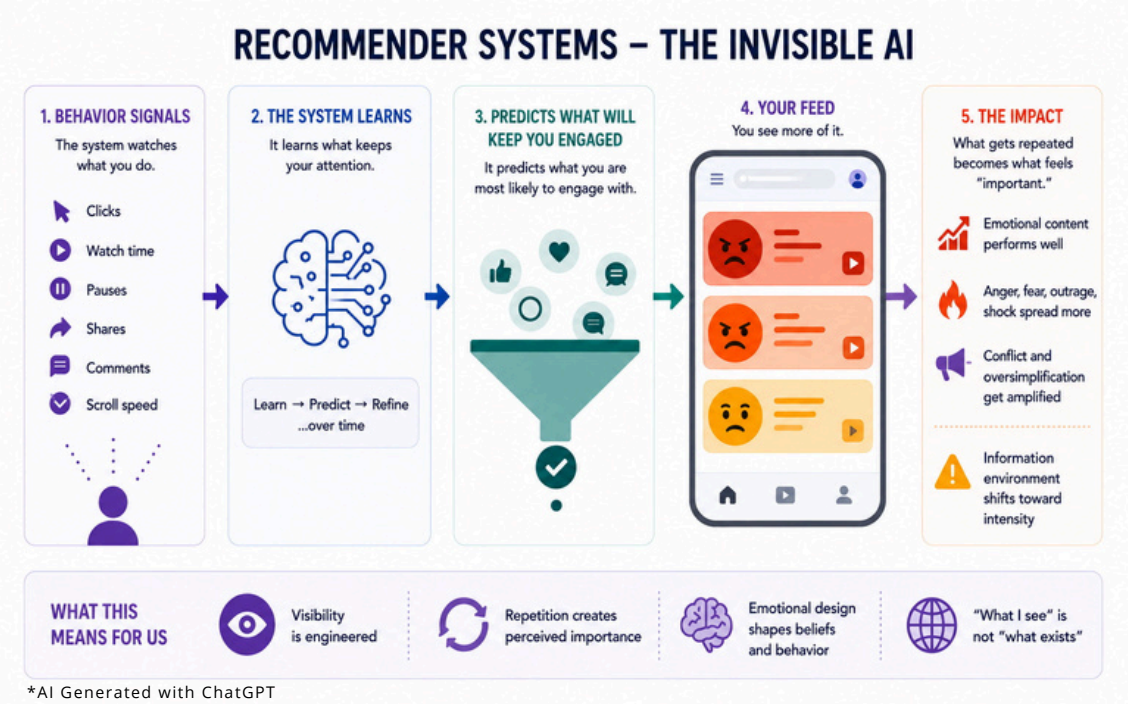
Generative AI creates new content by identifying patterns in large datasets and producing outputs that match a user request. It can generate text, images, video, audio, music, and code.

This makes it powerful for creativity and efficiency. It can help people draft content quickly, generate visuals, explore ideas, translate messages, or experiment with new formats without needing advanced technical skills.

But its limitations matter. It can create inaccurate or misleading content, reproduce stereotypes, generate unrealistic images, and make synthetic content appear more authentic than it is. It may also be used to create deepfakes, impersonation content, or large amounts of low-quality material. Users may also unknowingly share personal photos, documents, or sensitive information with platforms that store or reuse their data. Generative AI is best treated as a creative support tool—not as a source of truth. Users still need to verify information, check quality, and think critically about how outputs are used and shared.

This helps participants avoid two extremes: seeing generative AI as magical creativity or dismissing it completely. It can be useful—but its outputs still require human judgement.

E*xample: An image generator can create a realistic protest photo in seconds—even if the event never happened.*



*AI Generated with ChatGPT

9. Recommender systems - the invisible AI

For most people, the most influential AI is not a chatbot - it is the feed. Recommender systems shape what gets seen, repeated, and made "important." They do this by analyzing behavior signals: clicks, watch time, pauses, shares, comments, and even the speed of scrolling.

The core mechanism is simple: the system learns what keeps attention and predicts what will keep attention again. Over time, it refines this prediction for each user.

This matters because attention is not neutral. Emotional content - especially anger, fear, outrage, and shock - often performs well. So systems optimized for engagement can unintentionally amplify emotional extremes, conflict, and oversimplification. That does not mean every platform is "evil," but it does mean their incentives can push information environments toward intensity.

For MIL, the learning goal is not "hate the algorithm." It's to help participants understand that:

- visibility is engineered,
- repetition creates perceived importance,
- emotional design shapes beliefs and behavior,
- and "what I see" is not "what exists."

Example: *After watching two dramatic videos on one topic, your feed may quickly fill with more extreme versions - because the system learned you stayed longer when it felt intense.*

Data, privacy and safety

AI systems often collect more data than people realise, including prompts, uploaded files, search behaviour, device information, and usage patterns. This data may be used to improve systems, personalise experiences, or support advertising models.

For users, this can mean private information is stored, reused, or exposed in ways they did not expect. A practical habit is to avoid sharing highly sensitive personal, financial, medical, or work-related information in public AI tools.
Example: A user uploads a private work document for summarising without realising it may be stored or reused.

Example: A person shares personal medical information with a chatbot and assumes the conversation is fully private.

10. Personalization and profiling

Personalization means platforms tailor content to individuals. Profiling is how they do it: systems infer categories about users based on behavior data. This can include interests, fears, political leanings, vulnerabilities, or likely future actions.

Crucially, these profiles are not "truth." They are predictions. But they still have power because they shape what you see next, what ads you receive, and what messages reach you.

Personalization can be helpful (relevant content, accessibility, discovery). But it can also narrow perspective, reinforce habits, and create "separate realities" between users. Two people can live in very different media environments without realizing it, and this can deepen misunderstanding and polarization.

In facilitator language: personalization is not only about "you choosing content." It's also content choosing you—based on what the system predicts will keep you engaged. This reframing helps participants see agency as something to reclaim, not something automatically given.

Example: *Two friends search the same topic and get different results or recommendations, because the platform is tailoring what it thinks each person will click.*

11. Automation vs. autonomy

AI automates tasks. Autonomy implies independent judgement, responsibility, and moral agency—and AI does not have autonomy in that sense. It can generate, sort, recommend, translate, classify, and detect patterns, but it cannot understand consequences, care about harm, hold values, take responsibility, or be accountable.

This distinction matters because people often talk as if "AI decided" or "the algorithm wanted." That language can hide human responsibility—design choices, business incentives, policy gaps, and malicious use by individuals.

In MIL framing, this is where human agency becomes clear: humans build systems, choose what they optimise for, decide how tools are used, and remain responsible for harm—intentional or not.

Example: A platform may "recommend" harmful content, but humans set the goal (maximise engagement) and humans choose whether to share, report, or counter it.

Emerging development: agentic AI

Many AI tools still respond only when users give instructions.
But newer systems are being designed to act more independently by carrying out tasks across multiple steps.

These systems may:

- search for information
- compare options
- make recommendations
- complete tasks with less human input

This can increase convenience—but also raises new questions about control, responsibility, and oversight.

A possible consequence is that people may hand over decisions without fully understanding how choices are being made.

Example: An AI assistant may book travel, manage schedules, or complete online purchases with limited supervision.
Example: An automated customer service system may make decisions that are difficult to question or challenge.

12. Why AI improves unevenly

AI improves fast where training data is abundant, patterns are stable, and success is easy to measure. That's why image generation, translation, speech tools, and text drafting improve quickly.

But AI remains weaker where context changes fast, values and nuance matter, truth needs real-world grounding, and reasoning goes beyond pattern matching.

AI may create a convincing image but struggle to explain complex political issues accurately. It can sound confident about things it does not truly "know."

This unevenness matters: AI is powerful in specific ways, but limited in crucial human ways.

Example: AI can generate a realistic photo of a "new gadget," but it cannot reliably tell you whether that product actually exists or is being used to mislead you.

That concludes Annes A.
You do not need to be an AI expert. You need enough understanding to guide others with confidence, ask better questions, and keep human judgement at the centre.

*AI Generated with ChatGPT