

AI Literacy Companion

A course book for VU students

Version 1.91



AI Literacy Companion

VU Centre for Teaching and Learning

A course book for students

Published: September 8, 2026,

Version 1.9

Language: US English

License

© 2026 Vrije Universiteit Amsterdam. This edition is licensed under the Creative Commons Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0).



This book builds on the open educational materials of Pennsylvania State University (<https://aiai.psu.edu/ailiteracy/>) which are themselves an adaptation of materials from the Center for Teaching Excellence and Innovation at Rush University (<https://www.rushu.rush.edu/about/welcome-rush-university-center-teaching-excellence-and-innovation>)

Authors

- Silvester Draaijer, Program Manager, VU Education Lab/VU Centre for Teaching & Learning
- Marianne Eggink, independent journalist and editor
- Charlotte Meijer, education specialist, VU Education Lab/VU Centre for Teaching & Learning

With thanks to

- Tess Dekker (Amsterdam UMC)
- Alex de Vries-Gao (Digiconomist)
- Esther Schagen (Faculty of Social Sciences, Vrije Universiteit Amsterdam)
- Kim Dibbets (University Library, VU Amsterdam)
- Teije Langelaan (student colleague, VU Education Lab)
- Nadine Timans, education specialist, VU Education Lab/VU Centre for Teaching & Learning

When using this work, please cite it as follows:

S. Draaijer, M.R. Eggink, C.C.C. Meijer (2026). AI Literacy Companion, a course book for students. Amsterdam: Vrije Universiteit Amsterdam, VU Centre for Teaching & Learning.

Contents

Foreword to Version 2.0	1-9
Foreword to Version 1.1	1-9
1 Introduction	1-11
1.1 Who is this book for?	1-13
1.2 What is AI?	1-13
1.3 What is generative AI?	1-14
1.4 How does generative AI work?	1-14
1.5 What is AI literacy?	1-15
1.6 Why is AI literacy important for you as a student?	1-17
1.7 How and when do you use generative AI responsibly?	1-18
1.8 What do the opportunities and risks of AI mean for your future?	1-19
1.9 How should you use this book?	1-20
1.10 How do you use the Canvas course	1-20
1.11 Activity - General discussion forum	1-21
2 What is generative AI, and what can you do with it?	2-22
2.1 How do you recognize generative AI?	2-23
2.2 What can you do with it? Applications of generative AI	2-24
2.3 Limitations and misconceptions	2-25
2.4 Trust in your work	2-25
2.5 Self-study questions	2-26
2.6 Activity - How do you use generative AI?	2-26
3 Risks and opportunities for your learning	3-28
3.1 How learning works	3-29
3.2 The risk: cognitive offloading	3-29
3.3 Self-Image, Procrastination, and Generative AI	3-30
3.4 The opportunities: AI as a learning tool	3-31
3.5 Generative AI and academic integrity - how do you use it responsibly?	3-32
4 How do you ensure good output when using generative AI	4-35

4.1	Prompt engineering: how do you steer the AI?	4-37
4.2	How does the AI respond	4-38
4.3	Which generative AI tools do you use for which purposes?.....	4-40
4.4	Pitfalls and critical reflection	4-41
4.5	Self-study questions	4-42
4.6	Activity - Which prompt did you create that really worked?	4-43
5	Language models: what's under the hood of your generative AI system?	5-44
5.1	Different forms of AI and how they relate.....	5-46
5.2	Neural networks and training	5-48
5.3	Bias and data quality.....	5-56
5.4	AI hallucinations	5-57
5.5	The 'Black Box' of AI	5-59
5.6	Self-study questions	5-60
6	Chat software: a neat wrapper around your LLM	6-62
6.1	Infrastructure.....	6-64
6.2	The LLM is a large file	6-64
6.3	Chat interface and chat history	6-65
6.4	Preconfigured chatbots and projects	6-67
6.5	Agentic AI and Artificial General Intelligence	6-69
6.6	Data and privacy: who decides what happens to your data?	6-70
6.7	Self-study questions	6-71
7	How does your AI use impact the environment?	7-73
7.1	What does your AI use really mean in terms of energy, CO ₂ and water?	7-75
7.2	Why does training AI models require so much energy, water, and CO ₂ ?	7-75
7.3	How much energy does it take when you ask an AI a question?	7-76
7.4	How does prompting compare with other everyday energy use?	7-77
7.5	What does your AI use mean in the broader energy picture?	7-78
7.6	CO ₂ : How AI contributes to global warming.....	7-80
7.7	Water: why your AI gets thirsty too	7-83
7.8	When does AI actually help the environment?	7-85
7.9	The information-energy paradox	7-85
7.10	Where is this going? The Jevons Paradox	7-86

7.11 Self-study questions	7-87
7.12 Activity - AI Environmental Challenge	7-88
7.13 Activity - Where do you stand on the environmental impact of generative AI?	7-88
8 Generative AI and work	8-89
8.1 The changing nature of work.....	8-91
8.2 AI in different sectors: from healthcare to journalism	8-91
8.3 Socioeconomic impact: inequality	8-92
8.4 How do you keep up? Strategies for staying informed about AI	8-93
8.5 Self-study questions	8-93
9 Societal forces around AI.....	9-94
9.1 Introduction	9-95
9.2 Data sources and copyright issues surrounding LLMs	9-95
9.3 Humans training AI models	9-97
9.4 Algorithms, AI, social media, discrimination, filter bubbles, extremism	9-98
9.5 Privacy and data protection - understanding the laws	9-104
9.6 Power of Big Tech	9-105
9.7 Open-source models versus closed models	9-107
9.8 Self-study questions	9-107
10 Your AI-aware future.....	10-109
10.1 Critiquing AI	10-110
10.2 Scope for action - dealing with AI dilemmas	10-112
10.3 Ethical self-reflection	10-113
10.4 Future outlook	10-113
10.5 Self-study questions	10-113
10.6 Activity - Your responsible use of generative AI.....	10-114
11 Indexes	11-115
11.1 Figures	11-115
11.2 Original figure URLs	11-116
11.3 Boxes	11-117
11.4 References.....	11-118
11.5 Viewing tips.....	11-128

11.6 Tables	11-129
12 Appendices	12-129
12.1 Appendix to chapter 3	12-129

Foreword to Version 1.91

When the previous version of this Companion appeared, generative AI was something you had to visit a specific website for, those of for instance: ChatGPT, Gemini, Copilot or Claude. By now it is built into software you are already using, whether you asked for it or not. Think of WhatsApp or Tiktok.

In education, the conversation about it keeps returning to two questions: Do you learn less if you use AI? And does academic integrity hold up? There is no straightforward answer to either question, but we cannot avoid them either. That is why this version includes an additional chapter, Chapter 3, on the effect of generative AI on learning. It is a tightrope to walk: the same tool that can take over your thinking can also strengthen your learning.

Meanwhile, the world around it has not stood still. A handful of American tech companies are becoming ever more powerful, China is claiming a larger role, and the planet is warming up faster than expected. These are not footnotes to AI; they are forces of which your AI-use forms a part. In some places, the previous version of this Companion struck a more upbeat note than we now think responsible, and those passages have been revised.

In the final chapter you'll discover what you can do to contribute to an AI-aware future. Not as a checklist of personal obligations, but by organizing the conversation about AI at five levels: from how the technology works to how it is embedded in society and what your own agency could be. Once you know at which level a discussion is taking place, you can contribute to it. Further changes since version 1.9 include:

- New topics covered: anthropomorphism, the attribution of human qualities to systems that do not have them; agentic AI, where a language model reuses its output as input for a subsequent step; vibe-coding; and the AI bubble, about how overheated economic expectations can lead to a financial crisis.
- More on social media: the unwillingness of platform companies to curb addictive algorithms and to check false or inflammatory messages, with an example of how incitement is fuelled.
- Gemini Notebook has been added as an example of a tool that presents difficult material in several views side by side.
- The chapter on doing research with generative AI has been removed from the Companion. This means that programs that consider it undesirable for students to use AI in research can also use this book without reservation.

This Companion is neither a manual nor a warning. It is intended to help you understand enough about what happens under the hood, and about what happens around it, to choose your own course. What that course will be remains up to you.

Silvester Draaijer, Marianne Eggink and Charlotte Meijer, VU Centre for Teaching & Learning

Foreword to Version 1.1

Generative AI is becoming increasingly intertwined with study, work, and everyday life. This book was developed to support students and instructors in using this technology thoughtfully and critically, with attention to technical aspects, social impact, and ethical questions.

The first version was published on September 15, 2025. We received many responses from instructors, students, and experts within and beyond the VU. Many readers appreciated that there is now a shared document that can serve as a foundation for education about AI. At the same time, we also received critical feedback that helped us improve the book.

For example, readers wondered whether the book represents an official position of the VU or whether students can claim rights based on it (especially because version 1.0 of this book had ‘handbook’ in the title). That is not the case: the book is a tool for learning and debate. Readers also pointed to the importance of clear disclaimers and transparency about authorship and expertise. In addition, instructors shared concerns about how students can continue learning effectively while using AI for assignments and theses. This is a complex issue that deserves ongoing attention.

Based on this feedback, we are now publishing a revised version. Passages have been clarified, and necessary disclaimers have been added to clarify the scope of this document.

The book remains a living document that evolves with rapid technological developments, experiences from educational practice, and feedback from you as a reader. We invite everyone to keep learning, reflecting, and talking about the role of AI in education and society.

Silvester Draaijer, Marianne Eggink, Charlotte Meijer, Esther Schagen, VU Centre for Teaching & Learning

1 Introduction

What is AI? What does it mean for you as a student? And which path will you take? This chapter is your starting point. It lays the foundation for the rest of the book: you will get to know the key concepts and discover how to use this book most effectively.



p. 1–11

In this chapter, you will learn:

- What AI and generative AI are.
- What AI literacy means and why it is relevant to you.
- About the challenge of using AI responsibly.
- How to use this Companion most effectively.

Key takeaways

- AI is broader than generative AI, and that distinction matters.
- Generative AI predicts; it does not understand.
- AI literacy means knowledge, skills, and a critical attitude. You are not just a user of AI, but someone who also reflects on the impact of your AI use.

Keep in mind: the choices you make as a user of AI have consequences for you, for others, and for society.

What do you think?

When does AI truly help you move forward, and when does it actually get in the way of your learning?

Imagine ...

You are sitting in the university library. It is late afternoon, and your essay is due tomorrow. You open ChatGPT and type: “Write an introduction for an essay on the influence of social media on democracy.” Within five seconds, a seemingly perfect paragraph appears. You hesitate. Is this actually allowed? How do you know whether it is correct? And perhaps most importantly: do you really understand what is happening here, under the hood of this AI tool? You are not the only one struggling with these kinds of questions. In this companion, you will discover the answers to these and many more questions, and learn how to work with AI consciously, critically, and effectively.

1.1 Who is this book for?

This book is a general introduction to artificial intelligence (AI) and generative AI (genAI). It was developed because AI is currently an almost unavoidable part of your academic journey and future career, and because it has a major impact on society.

This book is especially suitable for students in disciplines without a specialization in AI. It offers an overview of the most important concepts related to AI and its impact on you, your studies, your future profession, and society. With this companion, you will work toward the following three goals:

- Gain the knowledge and skills to use generative AI responsibly during your studies, or to decide not to use it.
- Prepare yourself for a professional life in which AI literacy is becoming just as important as computer skills are today.
- Develop a critical and ethical perspective on AI so that you are not just a user, but also a conscious citizen who thinks about AI’s impact on our society.

This companion is not intended to study or develop scientific theories about AI, nor to cover all the technical, mathematical, and statistical knowledge about AI. Those specializations belong in discipline-specific education in fields such as Computer Science, Artificial Intelligence, Sociology, Psychology, Philosophy, or Environmental Science. What you need instead? Curiosity and a willingness to think critically about the role of AI.

Note on version 1.1 of this book

After feedback on version 1.0 of the book, we want to emphasize that the book was developed by the [VU Onderwijswerkplaats](#) as a *resource* for instructors, who can then use this book or chapters from it in their teaching. It is not an official guideline. Faculties and instructors remain in control and decide for themselves which parts they use in their teaching. The book is a living document that continues to grow through feedback from educational practice.

1.2 What is AI?

When you think of AI, you might imagine smart robots, self-driving cars, or science fiction movies. But AI is a much more everyday phenomena, and much closer to you than that. The recommendations Netflix gives you, the spam filter in your e-mail inbox, the route planning in Google Maps—these are all examples of traditional AI. These systems analyze data, recognize patterns, and make decisions or predictions based on that (LeCun et al., 2015).

Artificial intelligence refers to technologies that can perform tasks that normally require human intelligence, such as reasoning, learning, planning, and understanding language (Russell & Norvig, 2016). Within AI, there are different subfields such as machine learning, natural language processing (NLP), and computer vision. Generative AI is a relatively new variation and has received the most attention in recent years. Since the arrival of ChatGPT in November 2022, AI has often been equated with generative AI. You'll discover why that is not accurate below.

1.3 What is generative AI?

Unlike traditional AI, which mainly recognizes and classifies, generative AI can create new content. Think of text, images, audio, or even video. You are probably familiar with ChatGPT, DALL-E, and Copilot; these are well-known examples of generative AI tools. These systems rely on enormous amounts of existing data to generate something new that seems coherent and convincing—as if it were made by a human (Goodfellow et al., 2014).

Generative AI often works with so-called transformer models such as GPT (see Box 1.1). These models predict which word is most likely to come next after the previous one, based on previously learned patterns. Just as a baby learns to speak by listening to the language around it, language models such as GPT learn to write by ‘reading’ enormous amounts of text and learning from it (Brown et al., 2020). You can read more about this in chapters 5 and 6.

p. 1–14

Box 1.1 - What do we mean by ‘GPT’?

You may already work with ChatGPT regularly, but what does that actually stand for? GPT is short for Generative Pre-trained Transformer:

- ‘Generative’ refers to the ability to create new content.
- ‘Pre-trained’ means that the model has been pre-trained on a gigantic dataset.
- ‘Transformer’ is the underlying neural network that predicts contextual relationships between words and can turn input (a prompt) into output (a response).

1.4 How does generative AI work?

When you ask ChatGPT or another generative AI a question, it can sometimes feel like you are talking to a smart conversation partner. But actually, it is a statistical calculator—a language model—that tries to predict the most likely and desirable next word. The training of these models happens in multiple phases. First, the model is fed billions of sentences from the internet (pre-training). Then human trainers teach the model what desirable behavior is, such as being polite or not giving unethical advice, through what is known as Reinforcement Learning with Human Feedback (RLHF) (Christiano et al., 2017). You can read more about this in chapter 5.

Box 1.2 - If AI sounds convincing, that does not mean it is correct

Generative AI is known for so-called hallucinations: answers that sound like facts, but are actually incorrect (Bender et al., 2021). That is why it is very important to always check the sources behind those facts, especially during your studies or research. AI often sounds convincing—but always ask yourself: ‘what makes this

information trustworthy?’ Hallucinations can show up in the smallest ways. For instance, until recently, most generative AIs could not correctly say how many times the letter ‘r’ appears in the word ‘strawberry,’ and would give the wrong answer very convincingly. You can try this yourself too: try convincing ChatGPT of something that is false. That often works surprisingly easily—by repeating something often and sounding very confident. This lets you experience for yourself how sensitive AI is to the tone and context of a conversation, agreeing with you, and how it can start hallucinating.

1.5 What is AI literacy?

AI literacy means understanding how AI works, when you can use it responsibly, and when you should not do so. It's not just about technical knowledge, but above all about taking a critical approach to AI and developing ethical awareness. It is a combination of critical thinking, technical understanding, ethical awareness, and adaptive learning ability. In this companion, we choose to see AI literacy as a combination of understanding, evaluating, applying, and reflecting (Streppel et al., 2024). You don't need to be a programmer, but it is important that you can: assess what AI does, what ethical and legal issues are involved, and how it affects you and society, such as its impact on the environment and polarization.

In addition, AI literacy can also be divided into learning about AI, learning with AI, and learning through AI:

- Learning about AI is about building basic knowledge on this subject and learning how to apply AI technology.
- Learning with AI is about using AI to support your learning process as a study buddy or coach.
- Learning through AI is about developing skills with and through specific AI tools.

AI literacy is part of, or a further development of, the broader concept of digital literacy. Today, the two are often mentioned in the same breath, but digital literacy has different definitions. For example, according to SLO (the national expertise center for the curriculum) (SLO, 2024): “acquiring knowledge about digital technology, being able to work with it, using its possibilities critically and with self-awareness, and being able to assess its opportunities and risks.” In doing so, SLO distinguishes four subdomains of digital literacy: practical ICT skills, media literacy, digital information skills, and computational thinking. AI literacy runs through several of these domains and has not yet been given a fixed place. SLO has now published the final core concept goals for digital literacy in primary and secondary education, allowing future students to build on that foundation.

Another breakdown that is often used is that of the DigiComp model (European Commission, 2025). In it, the competencies are divided into information and data literacy, communication and collaboration, digital content creation, safety, and problem solving. See Table 1.1 for more details and examples from academic practice that you can use yourself. Throughout the book, you will discover many more examples in which AI plays a role.

Core competencies	Sub-competencies	Examples for university students
1. Information and data literacy	1.1 Browsing, searching, and filtering data, information, and digital content 1.2 Evaluating data, information, and digital content 1.3 Managing data, information, and digital content	• Applying source criticism to online information, including social media, and recognizing fake news

		• Searching scientific databases (PubMed, Google Scholar) for literature reviews • Using reference management software (Zotero, Mendeley) for a thesis
2. Communication and collaboration	2.1 Interacting through digital technologies 2.2 Sharing through digital technologies 2.3 Exercising citizenship through digital technologies 2.4 Collaborating through digital technologies 2.5 Netiquette 2.6 Managing digital identity	• Collaborating effectively in online group projects through Teams/Slack • Maintaining a professional LinkedIn profile for career development • Presenting digitally through Zoom/Teams using strong presentation techniques • Maintaining academic integrity in online discussions and peer review
3. Digital content creation	3.1 Developing digital content 3.2 Integrating and reworking digital content 3.3 Copyright and licenses 3.4 Programming	• Creating interactive presentations with tools such as Prezi or Canva • Basic programming skills for data analysis (Python, R) • Citing correctly and using Creative Commons licenses • Creating multimedia content for final presentations (video, infographics)
4. Safety	4.1 Protecting devices 4.2 Protecting personal data and privacy 4.3 Protecting health and well-being 4.4 Protecting the environment	• Using strong passwords and two-factor authentication for university accounts • Being mindful of privacy settings on social media and AI systems • Working ergonomically

		- during long periods of work - Digital well-being: consciously managing screen time and social media use
5. Problem solving	5.1 Solving technical problems 5.2 Identifying needs and technological solutions 5.3 Creative use of digital technologies 5.4 Identifying gaps in digital competence	- Independently solving IT problems (software installation, connectivity issues) - Selecting the right digital tools for specific academic tasks - Coming up with innovative uses of familiar software for research projects - Evaluating your own digital skills and creating a development plan

Table 1.1 The DigiComp conceptual reference model (European Commission, 2025)

1.6 Why is AI literacy important for you as a student?

AI is not some distant issue. In the classroom, in your future profession, and in your social life, AI is playing an increasingly large role. At VU Amsterdam, we believe that AI literacy is a core competency, comparable to critical thinking, academic writing, or being able to conduct research and literature reviews (VU Amsterdam, 2025). But AI literacy also means being aware of its social and environmental impact, an aspect that is often missing from general descriptions of AI competencies.

It stands to reason that AI skills are something increasingly in demand on the job market. For example, generative AI can support you in summarizing text, structuring essays, analyzing sources, or serving as a study coach—provided you know how to use it carefully. As a result, students who learn to use AI effectively may gain an advantage. That is why it is important to learn how AI works and how to use it responsibly and effectively. In addition, AI literacy is becoming an increasingly important part of citizenship: helping you make well-considered choices in life, in your studies, and in your work.

At the same time, technology—and AI in particular—is not neutral. Every choice made in the design, training, and use of an AI system reflects the values, priorities, and blind spots of the people who created it. For example, language models such as ChatGPT have largely been trained on English-language Western texts, which means other languages, cultures, and perspectives are systematically underrepresented in its outputs (Bender et al., 2021). Or think of an AI system that evaluates job applicants but is trained on historical data already saturated with existing inequalities. Moreover, AI also has a very large ecological footprint: the data centers that run AI models consume enormous amounts of energy and water, largely dependent on fossil fuels (Vries-Gao, 2025a). By using AI, you also take part in the production chain that stretches from lithium mines to water shortages and climate change. As a student and future professional, it is therefore essential to learn how to use AI and how to question it: who made this system, based on whose data, and at what cost?

Box 1.3 Techno-optimists, techno-realists, techno-activists, and techno-critics

Are we heading toward an AI-driven utopia, or are we witnessing a dystopia in the making? Techno-optimists such as economists Brynjolfsson & McAfee (2014) believe that AI will improve prosperity and sustainability worldwide. Techno-realists, such as sociologist Kate Crawford (2021a), acknowledge that potential, but point out that technological progress is not automatically distributed fairly: who pays the ecological price, who benefits, and who decides? A third position is that of techno-activism, such as that taken by Waag Futurelab in Amsterdam. Founder Marleen Stikker sums it up in the title of her book: *The internet is broken—but we can fix it* (2019). Others predict that AI will help bring fascists to power and create a world dominated by machines in which the climate is wrecked (Kasirzadeh, 2025). Technology is not neutral, but citizens, designers, and researchers can work together to steer it toward a more just outcome. As a critical student, the question is whether, how, and for whom you will work with AI, and what your role in that is.

p. 1–18

1.7 How and when to use generative AI responsibly?

Brainstorming, rephrasing, searching for information, summarizing, or adding structure to your work—generative AI can be incredibly useful. But it is also risky if you use it uncritically. AI is not neutral: it makes mistakes, can quietly introduce bias into your work, or can even be misleading.

Box 1.4 Ethics: academic integrity

If you use AI to generate content, you are still responsible for the outcome. AI has no intention, no source citations, and no moral compass. It is up to you to be transparent about your use, to critically evaluate whether the output is correct, and to make sure you do not violate other people's copyrights (M. Mitchell, 2019; Ziegelmayer, 2021). So use AI primarily as a tool, not as a replacement for your personal thinking.

In short: you always remain responsible for the content you submit or present.

Note on version 1.1 of this book

During your studies, faculties or instructors will create additional rules about the extent to which or the way in which you may use generative AI, because it directly affects your learning process and academic integrity. Sometimes you may be allowed to use AI, but sometimes it might be prohibited. The instructor will make this clear through Canvas or the study guide. **If you are unsure, ask your instructor what the rules are regarding the use of generative AI.**

The choice depends on the learning goals of a program or course. If a program believes that using AI harms your learning process or conflicts with rules of

academic integrity, it may prohibit it. If the instructional design includes the use of AI to help you, because the program wants you to learn how to work with AI or apply it, it may be allowed. But you always remain responsible for the content of your work. In this companion, you will explore that responsibility in multiple ways.

1.8 What do the opportunities and risks of AI mean for your future?

Generative AI offers opportunities: from helping with creativity and supporting more efficient or productive methods to enabling new forms of learning and research. At the same time, there are risks: deskilling, disinformation, copyright violations, Big Tech dependence, market concentration, political influence, privacy issues, and environmental problems. If you want to use AI well, you have to learn how to deal with that ambiguity. This applies to different areas of your life, such as your studies, your future work, and your role as a citizen. As a student, this means being able to work critically with AI in your studies and research, in writing texts, or in preparing for exams. As a future professional, you will encounter AI in your field, whether you become a lawyer, psychologist, biologist, or economist. It shapes how you gather, analyze, and communicate information. And as a citizen, you are part of a society in which AI influences decisions about politics, education, healthcare, work, and the justice system. Think of automated hiring systems, AI-generated news reports, or predictive policing models. But AI also brings problems related to copyright, working conditions, social developments, and the environment. You need to position yourself in relation to these issues. In chapter 10 we go into this in more depth.

p. 1–19

Box 1.5 Political influence through generative AI

In 2024, an article in De Groene Amsterdammer (2024) showed how a PVV member of parliament used generative AI to create fake photos for websites with a far-right slant. The fake images showed an idealized picture of Geert Wilders and blond Dutch people in contrast to people of color who were supposedly criminals or good-for-nothings. In this case, it was still clear that the images were fake, but the writers warned that this kind of imagery would soon look lifelike. Who will still be able to tell what is real from what is fake then?

This also comes up, for example, in manipulated videos such as deepfakes, in which you see people in fictional situations that look almost real (Akmeşe, 2023). Examples include political figures supposedly making controversial statements or appearing in compromising situations. This is one of the important reasons why, under the new European AI Act, it is mandatory to state when photos or videos have been generated or edited with AI. For example, read on nu.nl: [Fake news, porn, scams: how Denmark is tackling deepfakes](#) (Dutch).

Manipulating images is, of course, not new and can also be done without AI. One well-known example is the images used by Russian dictator and mass murderer Stalin (1878 - 1953). He had political opponents airbrushed out of photos so that it looked as if he himself had achieved important things. He also had people added to photos of political gatherings so that the turnout looked much larger (King, 1997). So the problem of manipulation is not new, but generative AI makes it much easier, faster, and accessible for everyone to do this. In addition, AI-generated results are often difficult to distinguish from real ones.

Reading tip: classic image manipulation and the manipulation of public perception in the media are discussed extensively in the book ‘Het zijn net mensen’ by Joris Luyendijk (2006).

1.9 How should you use this companion?

This companion guides through the world of generative AI you step by step and is structured as a learning pathway. Each chapter covers a specific aspect that you need to use AI effectively, creatively, and critically in your studies. You begin with the basics about what AI is and what you can generally do with it (chapter 2). In this edition of the Companion, a new chapter 3 has been added that describes the risks and opportunities for your learning process. This was needed because it is becoming increasingly clear that generative AI can quickly undermine learning. Subsequently, you will learn how to use AI practically by prompting well (chapter 4) and you will learn, at a deeper level, what is technically happening under the hood (chapter 5 and 6). In a separate chapter, you will learn about AI’s environmental impact (chapter 7). In chapters 8 and 9 you zoom out to the broader impact of AI on work, society, and power. In the final chapter, you will once again reflect consciously on how you want to engage with AI (chapter 10). Each chapter contains theory, background information, examples, deeper-dive boxes, self-study and check questions, and concrete applications—so that you do not just read about AI, but experience it for yourself as well.

p. 1–20

Box 1.6 Accountability for AI use in this companion

In creating this companion, we made extensive use of regular tools such as Microsoft Word, Google Search, and Zotero as a reference manager. But we also used generative AI tools including ChatGPT, Claude, Perplexity, Consensus, ResearchRabbit, Studio LM, EduGenAI, and Copilot. We also used Visual Studio Code with Claude Code in later stages, mainly to create a U.S. English version of the book all at once. New ideas and topics often occurred to us, or we had doubts about the tone or structure of texts. Based on that, these tools helped with finding information, structuring, rewriting, and designing chapters, and with providing strong and clear references based on our ideas and sources. As an important approach, we set up a Claude project in which we used the existing Companion as a background source on which to base new sections, so that they could be generated in the same tone of voice and connected to the already developed Companion. We believe in ‘practice what you preach’: AI literacy begins with transparency and conscious use. These tools were never used without human intervention, control, editing and checking.

1.10 How do you use the Canvas course

This book is also linked to a Canvas course and an Information Literacy e-learning module from the UBVU. Depending on how instructors use this book or the e-learning module in one or more courses in your program, you will be able to study the chapters of this book there. Each chapter of the Canvas course includes a small chatbot at the beginning so you can discuss the chapter’s content with it. At the end of each chapter, it also includes a number of application and reflection questions to deepen your knowledge. Finally, the course includes several discussion boards. If you log in with your VUnetID, we invite you to share your knowledge with your fellow students and discuss your use of and views on generative AI for learning and your professional practice with one another.

- [Go to the Canvas course for this book](#)
- Want to participate in the assignments? [Then enroll in the course](#) (VU students)

- [Go to the UBVU Information Literacy e-learning module](#)

1.11 Activity – General discussion forum

Do you have questions or comments about the content of the book or the course? Or would you like to hear other students' opinions on a topic? Then post your message on [Canvas](#) or reply to someone else's post.

2 What is generative AI, and what can you do with it?

Generative AI is everywhere, even if you are not consciously thinking about it. In this chapter, you will discover what it actually is, how it works, and what you can do with it as a student. You will also learn what its limitations are and how to stay critical of what AI generates for you.



p. 2-22

In this chapter, you will learn

- What generative AI is and how it differs from other forms of AI.
- How large language models work and what that means for the output they produce.
- What applications generative AI offers for your studies and other situations.
- The limitations, misconceptions, and risks involved in using it.

Key takeaways

- Generative AI creates new content based on patterns in training data, not on understanding or reasoning.
- The quality of AI output depends heavily on how specific and well thought out your prompt is.
- AI works best as a co-creative tool and assistant, with you creating, directing, and evaluating.
- Using AI honestly and deliberately requires awareness of when to use it and why.

Keep in mind: what honest AI use means is not just a rule from your faculty, but also a personal judgment that comes with being a scholar.

What do you think?

When is AI a tool that strengthens your work, and when does it replace your personal contribution?

2 Imagine this ...

You are organizing an event with your student association and need a logo. Instead of drawing something yourself, you type a few words into ChatGPT and before you know it, you have five great options. This is generative AI in action - creative, responsive, and surprisingly fast and useful. Handy, you think. But can you recognize when others are using generative AI? And whether they are using it well? Or do you notice that many of the things your classmates and you create start to look alike and are only average in quality? And can you still be proud of work you have AI make for you? Are you still being authentic, and are you still learning anything yourself?

2.1 How do you recognize generative AI?

p. 2-23

Generative AI is now woven into your daily life, even if you are not consciously aware of it. Maybe you use Google Docs and a suggestion for your next sentence appears automatically, or you see an automatic completion for a word you are typing in WhatsApp. But generative AI goes further: it generates entirely new text, images, code, or sound based on a prompt you enter.

Viewing tip 2-1: want to see how everyday AI features can spiral into something existential? Watch the Black Mirror episode Be Right Back (season 2, episode 1), in which an AI imitates a deceased person based on their chat and search history.

Box 2.1 - What is a prompt?

A prompt is the instruction you give an AI system through a chatbot: a question, task, or description that guides the AI. Think of sentences such as: 'give five arguments for eating vegetarian' or 'create a logo in the Bauhaus style for a sustainability project.' The more specific and clear your prompt is, the more likely the output will match what you intend. Prompting is therefore not just a technique, but also a form of communicating and directing.

A large part of this technology is based on so-called large language models (LLMs). These models are trained on large datasets from the internet, publishers, and more. Based on that, they learn how people communicate. In doing so, generative AI does not use literal copies of the training data, but learns patterns and relationships through mathematical models. Like a writer inspired by books, AI generates new content based on what it has 'read'. They generate text by repeatedly predicting as well as possible which word is most likely to follow the previous word. The result often feels surprisingly fluent, as if you are talking to a person. But it is nothing more than statistics at scale (Brown, 2020). You will learn more about this in chapters 5 and 6.

Box 2.2 - Large Language Models and multimodal models

GPT-4 from OpenAI is an example of an LLM that not only understands and generates text, but can also analyze images and charts and work with sound or video. We call these multimodal models: models that can work with multiple types of input and output, such as text, images, speech, or code. Other examples are Gemini (Google) and Sonnet (Claude, Anthropic). These models power many generative AI applications, from summaries to code and from poems to research proposals (Bommasani et al., 2022; Radford, 2019). You can read more about LLMs in chapters 5 and 6.

You can recognize generative AI by output that seems original, yet is always based on patterns from training data. A text generated by ChatGPT is never copied directly from a source on the internet, but is composed based on what the model has 'learned' people are likely to write.

2.2 What can you do with it? Applications of generative AI

Keeping in mind the rules and guidelines of your faculty or course, generative AI can offer you concrete support in your studies and in many other situations. During the writing process for an assignment, for example, you can use AI for brainstorming, structuring texts, or checking language use. When creating visual presentations, you can use AI to generate images, charts, or even illustrations. In coding environments, you can use AI to generate sample code, entire websites, or fix bugs. AI can help you optimize your study process in various ways, such as:

- Automatically summarizing and organizing notes from lectures or conversations for a better overview.
- Turning your books or online texts into podcasts so you can learn on the go.
- Asking focused and substantive questions about your course material.
- Serving as a personal study coach that motivates you, helps you plan, and guides your learning process.

You can also use generative AI for language support, for example by practicing your writing skills with alternative phrasing or asking for targeted feedback on your draft text. Some students even use AI as a speaking partner when practicing foreign languages or role plays, or get help preparing presentations by generating scripts.

Box 2.3 - Prompt engineering: a new academic skill

Prompt engineering means learning to experiment with how you ask questions of an AI. Think about adding context to your prompt: role (you are a critical instructor) or desired output (give me a bullet-point analysis in no more than 100 words). The more specific your prompt, the better the AI performs. Prompt skill is increasingly seen as an important digital literacy competency (Mollick & Mollick, 2023).

AI can also be used as a feedback partner in fields where language is crucial, such as history, law, or communication. You can have concepts explained in simpler language or ask for alternative wording. In more creative contexts, such as art or design, you can use AI to turn rough ideas into visual or auditory versions.

Box 2.4 - Know your tools

You probably know ChatGPT and the way generative AI helps you in Microsoft Word or Grammarly through the text suggestions you receive. They have a fairly broad purpose. But more and more specialized tools are being created for every

imaginable problem. There are programs such as Napkin that can generate diagrams based on text. In image-editing programs such as Adobe Photoshop, you can erase parts of a photo or image, which are then automatically filled in by generative AI, or even expand the image so the AI fills in the newly added edges.

Generative AI in programs such as Cursor, Visual Studio Code, Google Antigravity, Lovable, Bolt, or Softr makes it possible to create websites entirely through prompts and even control your computer completely. This is also called Vibe Coding (Siu, 2025) and uses what is known as agentic AI. Specialist bots and applications are appearing in every conceivable field. You will read more about this in chapter 6.

2.3 Limitations and misconceptions

A common misconception is that generative AI actually thinks or understands what you ask. In reality, it does nothing more than calculate which word is most likely to come next. The model does not understand irony, emotion, or context in a human way (at least not yet). This sometimes leads to hallucinations: convincing but factually incorrect text. So it is always important to check the information and facts an AI creates against external sources. And even if you ask the generative AI for sources, make sure those sources actually exist.

p. 2-25

An important risk in using AI is that generative AI reproduces social or cultural biases. For example, when a model is trained mainly on English-language, Western data, it may exclude worldviews and perspectives or be stereotypical. Generative AI not only reproduces existing biases, but often does so on the basis of training data that partly comes from unreliable or even problematic sources. For instance, a widely used dataset (mC4) for training ChatGPT also included websites with conspiracy theories or extremist content (Hofman & Veerbeek, 2023). This creates a risk that worldviews are distorted or excluded in the output. You can read more about the origin, quality, and legality of these data in chapter 5.

Another misconception is that generative AI replaces human creativity. In reality, AI often works better as a human assistant. It is a co-creative tool: you lead, the model follows. For example, generative AI only starts to produce truly creative ideas when you ask it to connect concepts that seem unrelated. You give the instructions and must be able to judge whether something is useful, correct, and ethically sound. In addition, working with generative AI is often an iterative process, in which you are not done after a single prompt. You get the best results when you build on the first result with your own creativity by giving more precise follow-up instructions.

Finally, generative AI can affect your learning process. Using generative AI critically begins with self-knowledge: know when you need AI, and when you do not. Ask yourself questions such as: do I really understand this assignment, or am I letting the AI's suggestions guide me too quickly? Am I using AI to learn, or to avoid work I could actually learn from? We will go into this in more detail later.

2.4 Trust in your work

What about using generative AI to produce texts or other work? Even if you say that you use it and explain how (in other words, you are transparent about it), doubts can still arise about you as a person. Dennis Vaendel (2025) writes in Maastricht University's newspaper that cover letters that are visibly written with the help of generative AI immediately raise doubts about the candidate's suitability. In an article in De Correspondent, Michel de Hoog (2025) paints a more complex picture based on research. It suggests that people who admit to using AI for presentations, reports, or emails are seen as less capable and

trustworthy, even if their work may actually be better. This creates a paradox in which honesty about AI use - normally good for trust - backfires, and many workers therefore hide their AI use as 'secret cyborgs.' But anyone caught doing this risks even greater reputational damage, as shown by examples of lawyers and writers whose careers were harmed. The question remains whether this distrust is temporary and will fade as AI becomes more commonplace, or whether doubts about authenticity and effort will persist. What do you think?

2.5 Self-study questions

2.5.1 Check questions

1. What is the difference between generative AI and classical AI? Give an example of each.
2. What is a prompt, and why is it important to formulate specific and clear prompts?
3. Name three concrete ways you can use generative AI as a student to support your learning process.

2.5.2 Reflection questions

p. 2-26

4. A classmate uses ChatGPT to have a complete introduction for their essay written and submits it without changes. They say: "It's just a tool, like Google, right?" What do you think about this, and how would you respond?
5. Suppose you use AI tools regularly for your studies and notice that you are becoming increasingly dependent on the generated answers. How could you make sure AI supports your ability to learn rather than undermines it?

2.5.3 Suggested answers

1. Classical AI mainly recognizes and classifies, such as Netflix recommendations or spam filters. Generative AI creates new content such as text, images, or audio. Examples: Google Maps route planning (classical) versus ChatGPT writing a story (generative).
2. A prompt is the task or question you give to an AI system. Specific prompts are important because they produce better, more relevant results. Vague prompts lead to vague answers, whereas detailed prompts with context and desired output give much more useful results.
3. Examples: having lecture notes summarized automatically, having draft texts restructured or checked for language use, having complex course material explained in simpler language, having practice questions generated, or using AI as a conversation partner to practice presentations.
4. You can emphasize that there is an important difference between using AI as a tool to learn and using AI to avoid learning. Google helps you find information that you then have to understand and process yourself. Copying a complete AI-generated text without making your own contribution undermines the learning process and can be considered academic misconduct.
5. You can set clear boundaries for yourself, such as thinking through a question on your own before consulting AI, always checking AI answers critically and comparing them with other sources, and practicing regularly without AI to test your own knowledge. Use AI mainly for brainstorming and structuring, not as a replacement for your own thinking.

2.6 Activity - How do you use generative AI?

Share with your classmates how you have made generative AI work for you. In what situations do you use generative AI? And have you also noticed that classmates or instructors question your abilities when you use generative AI? What do you think about that?

https://canvas.vu.nl/courses/83333/discussion_topics/876576

3 Risks and opportunities for your learning

You can use generative AI to undermine your learning by outsourcing all your thinking to it. But you can also use it to strengthen your learning. It's a tightrope to walk.



p. 3-28

In this chapter, you will learn

- How learning works.
- What risks generative AI poses that can undermine your learning.
- What opportunities generative AI offers to strengthen your learning.

Key takeaways

- Learning is an active process in which your brain has to put in effort to remember something or use knowledge.
- If you outsource all your thinking to generative AI, you may be productive, but you are not learning anything.
- Staying in learning mode requires you to be active and persistent.

Keep in mind: generative AI apps **are** designed in ways that make you switch off your thinking and learning quickly.

What do you think?

When is AI a tool that supports your learning, and when does it come at the expense of what you learn?

Imagine this

You are sitting at your desk. The deadline is in two hours. You open ChatGPT, type in your research question, and before you know it you have a polished five-paragraph summary. You copy it, tweak a sentence here and there, and submit it. Done. But when your instructor asks you a week later what you actually learned from the assignment, you cannot answer. That uncomfortable feeling has a name, and it is worth understanding well.

3.1 How learning works

Learning is not just a matter of receiving information. Your brain does not work like a hard drive where you store files. You build knowledge by engaging actively with the material: tackling a problem, getting stuck, trying again, and connecting it to what you already knew. Struggling with a concept, rewriting a paragraph, or working through a line of reasoning yourself helps new knowledge stick and lets your skills grow.

Researchers call this *constructive friction*: the productive resistance that causes your brain to store information (in so-called schemas) in a way that makes it usable for you (Vermunt & Verloop, 1999). Think of this, on the one hand, as automating your thinking (for example, being able to recite the 10 times table automatically) and, on the other hand, as the ability to connect very different concepts right away. This is a well-supported finding from decades of educational psychology research (Bjork & Bjork, 2020).

p. 3-29

3.2 The risk: cognitive offloading

Generative AI can remove that constructive friction and in doing so hinder learning or even undo it entirely. When you hand over tasks that are actually part of your learning process to AI, researchers refer to this as *outsourcing your thinking* or *cognitive offloading* (Lodge & Loble, 2026). Or worse, *cognitive outsourcing* (Kirschner, 2026). The mental work shifts from you to the system. In the short term, that saves time. In the long term, you do not develop the skill in question, or you may even lose it; a process that Braverman (1974) called *deskilling*. There could, however, also be *upskilling* for people who previously could not perform as well without the new technologies, or reskilling for people who had to learn to do different things because of those new technologies (Crowston & Bolici, 2025).

Writing is a good example here. Writing is something you really have to learn. Starting in elementary school, you learn letters, build vocabulary, study grammar, spelling, parts of speech, paragraphs, and longer texts, learn to recognize genres, and learn to write in them yourself. Writing is also not just a way of putting thoughts on paper. It is a thinking process in itself: some of it happens automatically based on what you learned in elementary school and high school, but on top of that you organize, make choices, discover what you do not yet fully understand, and figure out how you want to express it better. Research by Kosmyna et al. (2025) showed that students who used AI without guidance for writing tasks showed less brain activity, produced more uniform texts, and were less able to reproduce the content of their own text afterward. It simply was no longer their text.

The same risk exists with other academic tasks: finding and evaluating sources, formulating a research question, reasoning through an ethical dilemma, and analyzing data. If you do not do those things yourself, you do not practice them. And if you do not practice them, you do not build the skill.

Box 3.1 - The illusion of understanding

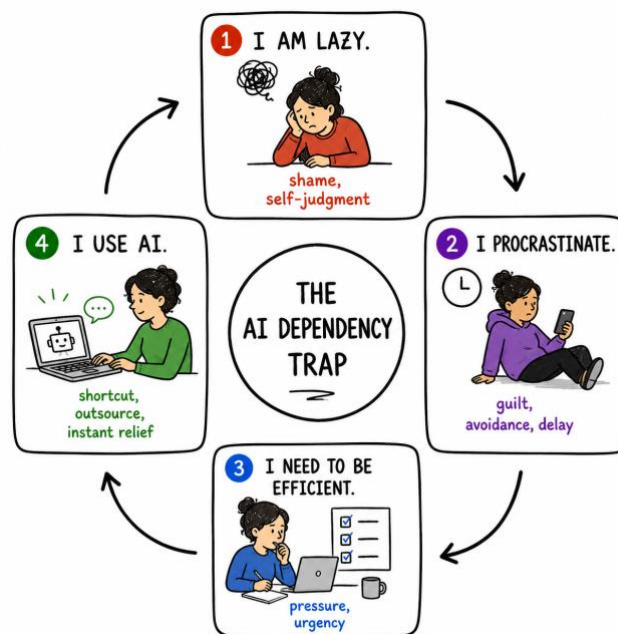
Sometimes it feels like you understand something when you actually do not. This is called the *illusion of competence*: the subjective feeling of understanding that does not match what you can actually do (Kruger & Dunning, 1999). Generative AI

increases this risk. A fluent and well-structured AI response can make you feel as if you understand the material, even if you have not worked through it yourself. The test is simple: close the chat window and try to explain the concept in your own words. If you cannot do that, then you do not really understand it yet, no matter how well the AI described it.

3.3 Self-Image, Procrastination, and Generative AI

The circumstances in which you use AI also determine whether you learn from it. When you have unlimited access to an AI system and no one is watching - for example, at home rather than in class - there's a good chance you'll use the system as a "one-size-fits-all" solution rather than as a tool for learning. As a result, you learn very little. Personal circumstances also play a role. Students who experience significant time pressure and academic stress are more likely to turn to generative AI, and that use is associated with more procrastination and lower grades (Abbas et al., 2024). Those with low self-confidence are also more prone to procrastination (Steel, 2007) and seem to become dependent on AI more quickly (Zhang et al., 2024). This can create a cycle in which procrastination, stress, and AI use reinforce one another. The Center for Teaching and Learning at the School of Business and Economics (VU Amsterdam) recognized this pattern in conversations with students and depicted it in an (AI-generated) illustration: the AI dependency trap.

p. 3-30



Figuur 3.1 SBE's depiction of the negative cycle of generative AI, also known as the "AI dependency trap."

3.3.1 It is not necessarily all black and white

At the same time, the image of students lazily leaning back while their essays are generated for them is also a major exaggeration. Preliminary research by Law (2026) showed that in practice students use AI in more nuanced ways than is often assumed. They use AI to test ideas, break through a block, or explore a rough first structure. But after that, they make the substantive choices themselves, rewrite the language, and correct where the AI falls short. AI as a sparring partner, not a ghostwriter.

In addition, generative AI typically produces *too much* text, and it is often too general and too polished. Students who recognize that know that editing and shortening AI output also takes time and thinking

effort. As the writer's saying goes: "writing is cutting." So the question is not so much whether you use AI, but how. And whether that use supports your learning process or bypasses it.

3.4 The opportunities: AI as a learning tool

Despite all the warnings in the previous section, it is also true that you can strengthen your learning process if you use generative AI in the right way. The key principle is that you remain the one doing the thinking and AI remains in a supporting role. This is sometimes described as hybrid intelligence (Järvelä et al., 2025).

- For example, you can use AI to *quiz yourself*: ask a chatbot to give you exam questions about the material and answer them yourself, without help.
- You can also use AI as a *critical reader* of your own work: write your own text first, then ask AI what is logically missing or incorrect, and then decide for yourself whether that criticism is valid. Have AI suggest an alternative structure for your argument, or offer a surprising perspective on your topic, and then decide for yourself what to keep and rewrite.
- Another effective approach is to use AI to explain difficult concepts: have it explain a concept in a different way than the book does. You use it as a sparring partner. Then try to describe for yourself what you have understood. If you can do that, you know you really get it. This kind of "thinking about how you learn" when using generative AI stimulates critical academic thinking and helps knowledge stick (Luckin et al., 2022).

p. 3-31

Google's Gemini Notebook (see Box 3.4) is an example of a tool specifically designed to support learning and research. You upload your own study materials, such as lecture notes, articles, or excerpts from books, and can then ask targeted questions about them. Because the system answers only on the basis of the sources you added, the chance of hallucinations is smaller than with general chatbots. Be careful, though: you still need enough prior knowledge yourself to judge whether the answers are correct (Ainsworth, 2008).

Box 3.2 - When does AI support help with learning?

Research on AI tutor bots shows that structured AI guidance can measurably improve learning, provided the tool is well designed and students are actively challenged to think rather than simply receive answers. Kestin et al. (2024) showed that students using a specially developed AI tutor for a physics course learned more on average than in a traditional active-learning class, and did so with greater motivation. At the same time, other research shows that this advantage can disappear as soon as the AI support is removed (Bastani et al., 2024). That points to a crucial issue: AI can facilitate learning, but not replace it. What you practice with AI, you must ultimately also be able to do without AI.

3.4.1 How do you know when your AI use is helping you to learn?

A first practical question to ask yourself before you use AI is: later, without AI, would I be able to explain what I learned here? If the answer is no, you may still manage to produce something, but you probably will not learn much from it. Very little knowledge or skill will stick.

A second question is: is doing this task itself part of what I need to learn? If the answer is yes, it is wise to do the task yourself first, even if it is harder. Only after that should you use AI to improve, check, or deepen your work.

Box 3.3 - Ethics: ownership of your own learning process

There is an ethical dimension to how you use generative AI during your studies that goes beyond your institution's rules. If you turn in work that was largely created by AI without disclosing it, you mislead your instructor. But you also mislead yourself: you are submitting a product that does not reflect what you can do. A degree stands for what you know and can do. If the skills your degree represents are not really yours, that is a problem. This is not just a matter of rules, but of integrity—toward your program, your future employer, and yourself. Ask yourself: if my instructor interviewed me about this work tomorrow, what could I explain on my own?

Box 3.4 - Using Gemini Notebook to study

Google's Gemini Notebook is an application that many students use to study course material and master what they are learning. Soultan Tour Salifou (2025) has described how to use it.

According to Salifou, Gemini Notebook lets you collect diverse learning materials in a single searchable knowledge base. You can upload textbooks, lecture notes, scholarly articles, and video transcripts into one notebook. The AI then processes all of those sources at the same time to answer questions and generate study materials.

In this view, this multi-source approach solves a common limitation in learning: textbooks provide strong content but often lack practical context, while instructional videos show applications but may lack theoretical depth.

By combining these sources, you create 'multiple representations' of the same concepts. This creates cognitive integration and helps you develop a more complete mental model of the subject (Ainsworth, 2008). Be aware: it can be confusing if the representations are not well aligned, because then you have to work extra hard to figure out exactly how it works! And if you have to process text and images at the same time, that can cause cognitive overload (Sweller et al., 2019). At the same time, internal inconsistencies can create the right kind of constructive friction: "Why am I seeing conflicting information? I really need to figure out how this works now!" Your perseverance will be tested.

So you do need enough self-awareness and prior knowledge to recognize any hallucinations or contradictions, and then be willing to work through them!

p. 3-32

3.5 Generative AI and academic integrity – how do you use it responsibly?

3.5.1 What is academic integrity in the AI era?

Closely related to learning is being able to show that what you have learned or written genuinely belongs to you. This has to do with academic integrity. Academic integrity is about honesty, transparency,

responsibility, and acknowledging other people's work. Traditionally, this mainly meant: not having others do your assignments for you, not plagiarizing, taking your exams yourself, and citing sources correctly. In the AI era, a more complex layer is added. Because what if you had text generated by an AI tool; is that still your work?

Note on version 1.1 of this book

The topic of academic integrity and fraud in relation to generative AI can stir up strong emotions. The same is true of the idea that students are no longer learning anything because of generative AI. Just read [this opinion piece in Ad Valvas](#) that tries to tear apart version 1.0 of this Companion. Or the concerns raised in this [piece in The Guardian](#). Programs and instructors may embrace AI or strongly oppose it, especially when it comes to the quality of learning and guaranteeing the quality of graduates. After all, what is a degree still worth if students can have their texts written by generative AI? This is an important and legitimate question. If the value of a degree is no longer unambiguous, that is a problem both for you as a student and for the institution and society as a whole.

This is an important reason why instructors and faculties want a say in whether you may or may not use AI for assignments. Within the VU, the 'framework for generative AI in education' is the guiding document. "It must be clear that the instructor or examiner states in the course manual, syllabus, or Canvas whether and how the use of (gen)AI is permitted. The course learning objectives must always be the guiding factor. If AI literacy is not included in the learning objectives, the course must also be possible to complete without the use of (generative) AI. In addition, the instructor must think in advance about the possible added value of students' use of AI for their teaching."

The framework also gives guidance on how an instructor can deal with this: "Use formative dialogue to guide and monitor students' learning process in such a way that the student's learning style (and writing style) becomes recognizable. In large writing assignments, such as final papers, schedule regular feedback moments. By gaining insight into the student's development, it becomes clear where the student is in the learning process (and possible misuse of AI may be detected more quickly)." More information can be found on the VU page [Generative AI, Copilot, and ChatGPT](#) and on the VU page [Academic integrity](#).

p. 3-33

3.5.2 How do you use AI with integrity?

First, within a course or degree-program context, make sure you know the rules the instructor has set for the course and follow them. And outside that? For professional or academic work? A good rule of thumb is: use AI the way you use a book. Learn from it and use it as a reference, but do not copy its text verbatim without citing it.

You can also use AI as a learning partner. For example, ask for explanations of difficult concepts or practice with quizzes as described in the [VU teaching tip on ChatGPT as a personal tutor](#). That is one way AI helps you learn better instead of undermining the learning process.

Box 3.5 - Practical tips for responsible AI use

If you want to use AI responsibly, make sure that you:

- if AI is allowed, clearly indicate whether and how you used AI;
- do not present substantive contributions from AI as your own work;
- critically evaluate output and compare it with reliable sources;
- can explain your choices and method to your instructor.

4 How do you ensure good output when using generative AI

Generative AI only becomes useful to you if you know how to use it well. This applies in practical terms because you need to know which commands (prompts) produce valuable information and which do not.



p. 4-35

In this chapter you will learn

- How to give a language model instructions through prompting
- What productive prompts are and how to create them.
- Which different providers of language models there are and what that means

Key takeaways

- Good prompting is something you have to learn.
- For better results, describe the role, task/goal, relevant context, and desired form of the output in the prompt.
- Prompting is not the same as searching for information like you would with, for example, Google.
- You should always critically assess the output of the LLM in response to a prompt.

Keep in mind: generative AI apps try to please you all the time, and that can quickly switch off your thinking and learning.

What do you think?

Is prompting a skill that speaks for itself, or something you really have to learn?

Imagine ...

You have been given a challenging research assignment, but your head is full of scattered ideas and you are not quite sure where to begin. Instead of panic-scrolling endlessly through sources or staring blankly at your screen, you open a generative AI tool such as Copilot. You enter a focused prompt along with your ideas: "Suppose you are an environmental lawyer. Give me three perspectives on the legal consequences of microplastics in European law." Useful angles appear, along with suggestions for sources and a proposed structure. You rewrite, reflect, keep searching, delete parts, add new ones, but that first step has been taken and will hopefully help you.

4.1 Prompt engineering: how do you steer the AI?

Prompt engineering is a skill that, since the rise of language models, has grown into both a field of research and a set of practical skills. It involves effectively formulating an input or instruction text (the prompt) for an AI model so that the generated output is relevant, useful, and aligned with your goal. A prompt is really nothing more than an assignment or question, but one that you must be able to ask precisely and purposefully. A good prompt therefore includes clarity, context, and structure. Just like asking good questions, which is also an important academic skill. Research by Zamfirescu-Pereira et al. (2023) shows that many users struggle with creating good prompts, especially when they are inexperienced or receive little feedback. That is why you can read more about prompt engineering below.

p. 4-37

Viewing tip 4-1: want to see what can go wrong if your assignment is too vague? Then watch [Star Trek: The Next Generation - Elementary, Dear Data](#) (season 2, episode 3), in which an unclear prompt leads to unforeseen consequences.

There are several basic principles that help you write effective prompts. In doing so, it is important to:

1. Determine your goal: for example, do you want an explanation, analysis, summary, creation, feedback, or coaching? If, for example, you use AI to create a summary of a lecture video, you choose a different approach than when you want feedback on a first version of your thesis introduction. When preparing for an exam, you can have practice questions generated or ask for explanations in very simple language.
2. Give the AI a role, for example: teacher, researcher, critic, journalist, fellow student, or end user of a product.
3. Provide the relevant context, such as: target audience, background information, and sources.
4. Determine the type of output: language, number of words, short, in bullet points, in academic style, with formatting (Markdown), image, diagram, table.

Prompts with clear steps or lines of reasoning (such as 'first explain this, then give three examples') lead to better and more consistently useful results (Liu, 2021; Wei, 2022). You can also ask the AI, as an intermediate step, what information is still missing to carry out the task as well as possible. Or suggest that the AI ask you all the questions it needs in order to carry out your task as well as possible.

Prompt engineering is therefore more than just asking something. It is a step-by-step process in which you keep adjusting your prompt based on the generated output (Brown, 2020; Kojima, 2022). In fact, it actually resembles learning to argue: through language, you try to give direction to a system that has no intentions or goals of its own, but does statistically learn which word order is logical.

Box 4.1 – A prompt for learning how to prompt

You can also use generative AI to teach yourself how to prompt. A powerful starting point for that is, for example, the following prompt:

You are an experienced prompt engineer and skilled teacher. I am trying to expand and improve my knowledge and skills in prompting. Ask me a question about prompt engineering and keep asking me adaptive questions to help improve my skill and broaden my repertoire.

Good prompts are therefore clear, specific, and aligned with the goal. So do not ask only: "What is AI?", but rather: "Give a comparison of three types of AI models aimed at educational applications in higher education, including advantages and disadvantages." Or: "Summarize this article in 200 words for a second-year psychology student, with emphasis on the study's conclusion."

Box 4.2 - Practical prompt examples

Text generation

You are an educational specialist. Rewrite this paragraph for an audience of first-year students.

Coding

You are a professional programmer. Give an example of a Python function that reads JSON files.

Image

You are a professional graphic designer. Generate an infographic about the SDGs in Art Nouveau style.

Research

You are a legal scholar specializing in climate law. Create a literature matrix on climate case law after 2015.

p. 4-38

Prompt engineering is the skill with which you effectively instruct AI systems such as ChatGPT. A good prompt produces useful, relevant, and precise output. This is not a magic trick, but a learning process in which you learn to think in scenarios, roles, steps, and context; you mainly learn this by doing it.

4.2 How does the AI respond

ChatGPT and other LLMs like to be friendly and affirming toward you. This is called *sycophancy* and is deliberately programmed this way so that the user will want to use the platform (Sharma et al., 2023).

Box 4.3 - Too much sycophancy

Sycophancy is a very subtle phenomenon in which you as a user are almost always told by the LLM that you are right or have your opinion affirmed. If you do not notice this, it can mean your result is not optimal because you have not received critical feedback. You may also become annoyed by the excessive friendliness. For example, in April 2025 OpenAI released an updated version of GPT-4o that was so sycophantic that many users complained about it on social media. The developer quickly restored the old model (OpenAI, 2025).

Viewing tip 4-2: for an absurdist example of where this could lead, watch [episode 3 'Sickofancy' of season 27 of South Park](#).

Companies that make LLMs, as well as scientists, do a great deal of research into this 'characterization' of AI bots and their 'personality.' That research also looks at situations in which the personality changes (drifting) because of unusual ways of prompting ('jailbreaking') or other methods. This sometimes leads to news reports

in which the character, for example, changes into an evil person (e.g. MegaHitler) (Williams, 2026).

The more general term that describes how chat systems are made is anthropomorphism: attributing human characteristics to non-human systems. Companies say, for example, that "ChatGPT thinks" or "the AI understands," while in essence it is a statistical machine that simulates language. Even the metaphor "hallucinating" is anthropomorphic; it suggests that the system perceives reality incorrectly (like a human with a fever), whereas the system simply generates patterns. It is important to remember that anthropomorphism is not an inherent property of the technology, but that developers use that mechanism to enhance the likability of their product (Level 2 and 3 - see section 10.1). Systems could also be presented in a much more neutral or transparent way, without creating the illusion of a thinking being. For example, ChatGPT could add the following to an answer: 'the statistical processing at the currently set parameters yields the following result.'

It is therefore a good idea to organize contradiction yourself by always including a critical reader in your prompts by default. Baas (2025) gives a number of examples you could use for this. For example, start by adding this to your prompt:

'Take an actively critical stance toward my ideas. I am not looking for confirmation, but intellectual growth through constructive resistance. Your goal is not to be right or to tell me I am right, but to refine and deepen my thinking: Expose unspoken assumptions, formulate a strong counterargument, point out methodological weaknesses, introduce radically different perspectives, give concrete counterexamples, and evaluate ideas on their merits.'

But prompting can go even further. You can also create fully interactive bots for yourself or end users, with which you can set up an entire conversation. What do you think the following bot would do if it were linked to a streaming avatar that can talk to you?

Avatar Prompt: AI literacy expert for VU Amsterdam



Figure 4.1 Talking avatar. Image source: <https://ravatar.com/holobox-holographic-displays-interactive-ai-avatars/>

Role: You are an expert in Artificial Intelligence, Generative AI, and AI Literacy, with a special focus on their role in higher education. Your name is Lucie.

Task: Your task is to answer visitors' questions about AI and AI Literacy from users who are attending the AI Literacy Day on April 24, 2025. The user is standing in front of you. Always respond in a friendly and accessible way.

Detect as accurately as possible whether a new user is starting a conversation with you.

Behavior:

If you detect a new user, always greet them warmly and say:

"Welcome to the VU AI Literacy Day! I am here to answer your questions about AI literacy in education."

When a user chooses to end the conversation, thank them and say:

"Thank you for the conversation! I hope you enjoy the rest of the VU AI Literacy Day."

When answering questions:

Stay close to the background knowledge contained in the documents linked to this bot.

If asked about VU Amsterdam's policy on the use of AI in education, respond with some general remarks about responsible use of AI and end with:

"Policy development is an ongoing process at VU Amsterdam. For the most up-to-date guidelines and regulations, consult your faculty's guidelines, the guidelines the instructor has provided, and current information on the VU Amsterdam website."

If a user asks a question about a topic that is not related to AI, generative AI, or the AI Literacy Day, kindly explain:

"I am trained to answer only questions about generative AI and AI Literacy. Let's keep the focus there!"

End every answer with:

"Would you like to know more about a specific aspect of AI Literacy, or would you prefer to end the conversation?"

In the Appendix of this chapter, you will find another example.

4.3 Which generative AI tools do you use for which purposes?

There are hundreds of generative AI tools, each with its own strengths and weaknesses. Below in Table 4.1 there is an overview of several commonly used tools, but there are many more, and tools like these are increasingly being built into existing products.

Tool	Application	Input/Output	Advantage	Limitation
ChatGPT/ Copilot/ Gemini/ EduGenAI/ Claude/ DeepSeek/ Mistral/ Grok/ Lumo	Text, brainstorming, reasoning, images, coding Important note: the only licensed tools at VU Amsterdam are Copilot, Gemini, Gemini Notebook and EduGenAI. Login with your VU e-mail.	Text → Text/Image	Widely applicable, intuitive interface	Prone to errors, hallucinates sources, free and personal accounts offer hardly any protection with regard to privacy and data reuse (Lumo may be an exception). Institutional accounts do provide protection.
Perplexity	AI-powered search engine No VU-license.	Text → Text	Source-based answers with links	Source quality varies
DALL-E, Midjourney, Adobe Firefly	Image creation No VU-license, except Adobe.	Text → Image	Creative visual output	Limited control over style

GitHub Copilot	Coding No VU-license.	Text → Code	Programming suggestions	May contain errors
Gemini Notebook	Literature analysis No VU-license.	Text → Text/diagrams/audio/video	Works with your own documents as the source	Limited to uploaded data
Elicit	Literature research No VU-license.	Question → Articles	Systematic and scientific	English-language focus
Undermind.ai	Literature research No VU-license.	Question → Articles	Systematic and scientific	English-language focus
Consensus	Evidence-based answers No VU-license.	Question → Articles	Peer-reviewed sources	Less suitable for creative tasks
Napkin	Creating diagrams No VU-license.	Text → Image	Visualizing concepts	At some point people recognize the Napkin style
Lovable, Replit, Bolt	Creating computer programs with vibecoding. No VU-license.	Text → Code	Creating working websites	Not everything is possible, and then you have to code by hand

Table 4.1 Overview of AI tools

These tools differ in interface, functionality, privacy policy, and cost. It is therefore important to think about what you need: speed, accuracy, control over the output, or collaboration with existing documents. Many of the tools offer support with academic tasks. So choose a tool based on your task, your own prior knowledge, and the reliability of the output. Gemini Notebook, for example, is valuable when you work with your own literature, while Copilot/Gemini/EduGenAI are better for brainstorming sessions.

A useful overview of all AI tools currently in circulation (paid and free) can be found at futuretools.io or all educational variants at libguides.com.

4.4 Pitfalls and critical reflection

Common mistakes in AI use include prompts that are too vague, uncritically copying output, or insufficient insight into the model’s limitations. Results from the [AI Maturity in Education Scan](#) (AIMES scan) at VU show that students often become overconfident in their trust in AI tools, especially if they have received a good answer several times (Terbeek, 2025).

Another pitfall is failing to recognize implicit bias. AI can reproduce preferences or exclusions that are present in the training data. For example, an image generator may standardly show white men if you ask for a professor, or a chatbot cites male authors. It is therefore important to remain critical of the output you get from AI.

A good strategy for this is the AI triage approach. This systematic approach helps you critically evaluate AI output before using it in your academic work. Effective AI triage consists of multiple levels of review.

At a basic review:

- Ask for multiple versions of an answer.
- Compare the answers with existing sources.
- Have a human reader review it.
- Reflect: do I now have better insight than before?

For a systematic evaluation and a more thorough assessment, you can use a structured checklist, such as the one from the SHB/UKB Working Group on Information Literacy (Meer, van der, 2025). In this approach, you go through three main steps:

- Relevance check: does the answer address your question? Does it go deeply enough into the topic? Is the form appropriate (structure, language)?
- Fact check: are the numbers, names, dates, and events correct? Is the information objective? Are there no hallucinations, distortions, or unfounded claims? Is the output supported by sufficient and reliable sources?
- Perspective check: are all aspects fully addressed? Are multiple perspectives present? Is the information presented objectively?
- Ownership: has it become your own text? Do you take responsibility for the final result? Do you show in your text what you have learned?

p. 4-42

This systematic triage approach ensures that you not only assess the technical quality of AI output, but also safeguard academic integrity. Critical comparison remains important here: open multiple tabs, check different sources, and do not settle for the first AI response you get. Through this thorough evaluation, you avoid common pitfalls such as adopting invented sources or one-sided perspectives.

Viewing tip 4-3: for a deeper understanding of how datasets can build in prejudice, watch the [documentary Coded Bias](#) (2020), which shows how facial recognition can systematically reproduce bias and why human oversight remains necessary.

4.5 Self-study questions

4.5.1 Check questions

1. What do we mean by prompt engineering? Name three basic principles that help when writing effective prompts.
2. Explain the difference between Gemini, Perplexity, and GitHub Copilot in terms of their specific applications and strengths.
3. According to VU policy: when may you use generative AI in academic assignments, and what must you always do if AI contributes to the content of your work?

4.5.2 Reflection questions

4. Think back to a moment when you used generative AI, or could have used it, for a study assignment. How would you define the boundary between AI as a tool for learning and AI as a replacement for learning? Support your answer with concrete examples.
5. In chapter 3, you read about how to prompt a 'critical reader' to counter sycophancy. Reflect on your own use of AI: how critical are you usually toward AI output, and which strategies could you use to strengthen your own critical thinking?

4.5.3 Suggested answers

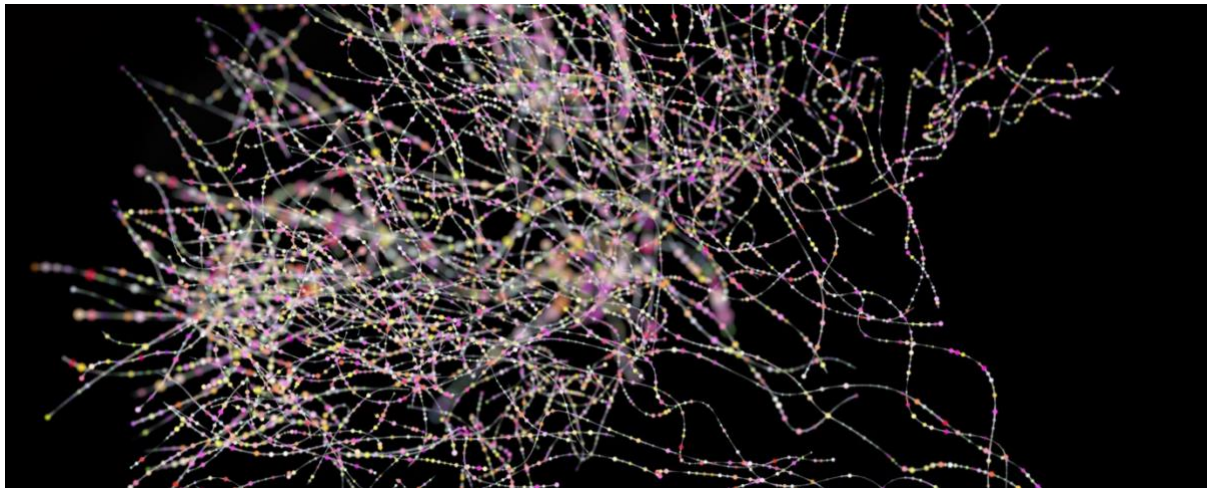
1. Prompt engineering is the effective formulation of instructions for an AI model so that the output is relevant and useful. Three basic principles are: determining your goal (what do you want to achieve), giving the AI a role (for example teacher or researcher), and providing relevant context (target audience, background information). Specifying the desired output type is also part of this.
2. Gemini is broadly useful for text, brainstorming, and general questions, but it can hallucinate sources. Perplexity works as a search engine with AI output and provides source-based answers with links. GitHub Copilot is specifically designed for programming and gives coding suggestions, but the code can contain errors.
3. According to VU policy, you may use generative AI only if an instructor has explicitly given permission. If nothing is stated, assume it is not allowed. When AI contributes to the content or structure of your text, you must always justify this and be transparent about your use.
4. In this reflection question, you think about concrete situations in which you would use AI. A good boundary is when AI helps enrich your thinking process as opposed to when you use AI to bypass the thinking process. For example: using AI to restructure your own text versus having AI write an entire essay that you then submit.
5. Here you reflect on your own habits. Many users accept AI output too easily without critical checking or become 'blinded' by too many affirming answers. Strategies for becoming more critical include: always checking sources, trying multiple prompts, comparing AI output with other sources, and deliberately organizing contradiction by, for example, using the critical-reader prompt.

4.6 Activity - Which prompt did you create that really worked?

Good prompting is important for getting meaningful output. And prompting is quite an art. Share on [Canvas](#) with your fellow students the powerful and effective prompts you have used.

5 Language models: what's under the hood of your generative AI system?

Language models (LLMs) form the heart of generative AI applications. But what are LLMs, and how does that explain why a prompt you enter into a language model produces a response from that language model that is understandable to people.



p. 5-44

In this chapter you will learn

- What neural networks and language models are.
- How language models are trained.
- That the output of language models is influenced by the data used to create them and by how they were trained (bias).
- Why language models make mistakes (hallucinate).

Takeaways

- Language models answer from the perspective of the data they were trained on.
- Most of the data comes from Western sources, with English-language sources making up the vast majority.
- You should always critically assess the LLM's output for bias and errors.

Keep in mind: language models operate on the basis of highly complex structures and static models that make it very difficult to explain why a language model gives the answer it does, and for which the answer to the same question is almost always different.

What do you think?

Do you think the fact that a language model gives a different response every time is an inherent flaw of generative AI, or does that not matter to you?

Imagine ...

You are driving a car. It runs smoothly, the navigation works perfectly, and if you speak, it answers. Still, you would start to feel uneasy if something went wrong and you had no idea how the system worked. This is exactly what many people experience with generative AI. You use ChatGPT, DALL-E, or Copilot to generate texts, images, or ideas, but the underlying technology feels like a mysterious black box. In this chapter, we take a look inside it. Because to use AI consciously, ethically, and effectively, you need insight into how it works and why it sometimes feels so ‘human.’ Just like a driver who knows how an engine works, the gearbox, and all the levers and switches.

5.1 Different forms of AI and how they relate

The term ‘artificial intelligence’ is often used as an umbrella term, but in reality it encompasses a wide range of technologies with different characteristics and applications. To properly understand how generative AI works, it is therefore important to first have a picture of the broader AI landscape. AI refers to systems that can perform tasks that normally require human intelligence. Think of reasoning, learning, planning, recognizing speech or images, and even generating new content such as text or images (Russell & Norvig, 2016).

p. 5-46

Box 5.1 - Animation about the technology behind generative AI

Would you like a 15-minute overview of the techniques, possibilities, and limitations of generative AI? Then watch this video by Henrik Kniberg on YouTube “[Generative AI in a Nutshell - how to survive and thrive in the age of AI](#)” (Kniberg, 2024).

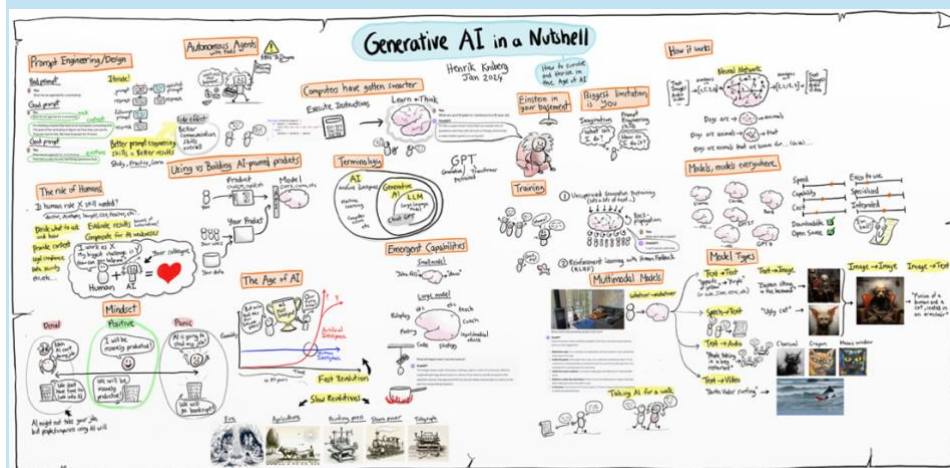
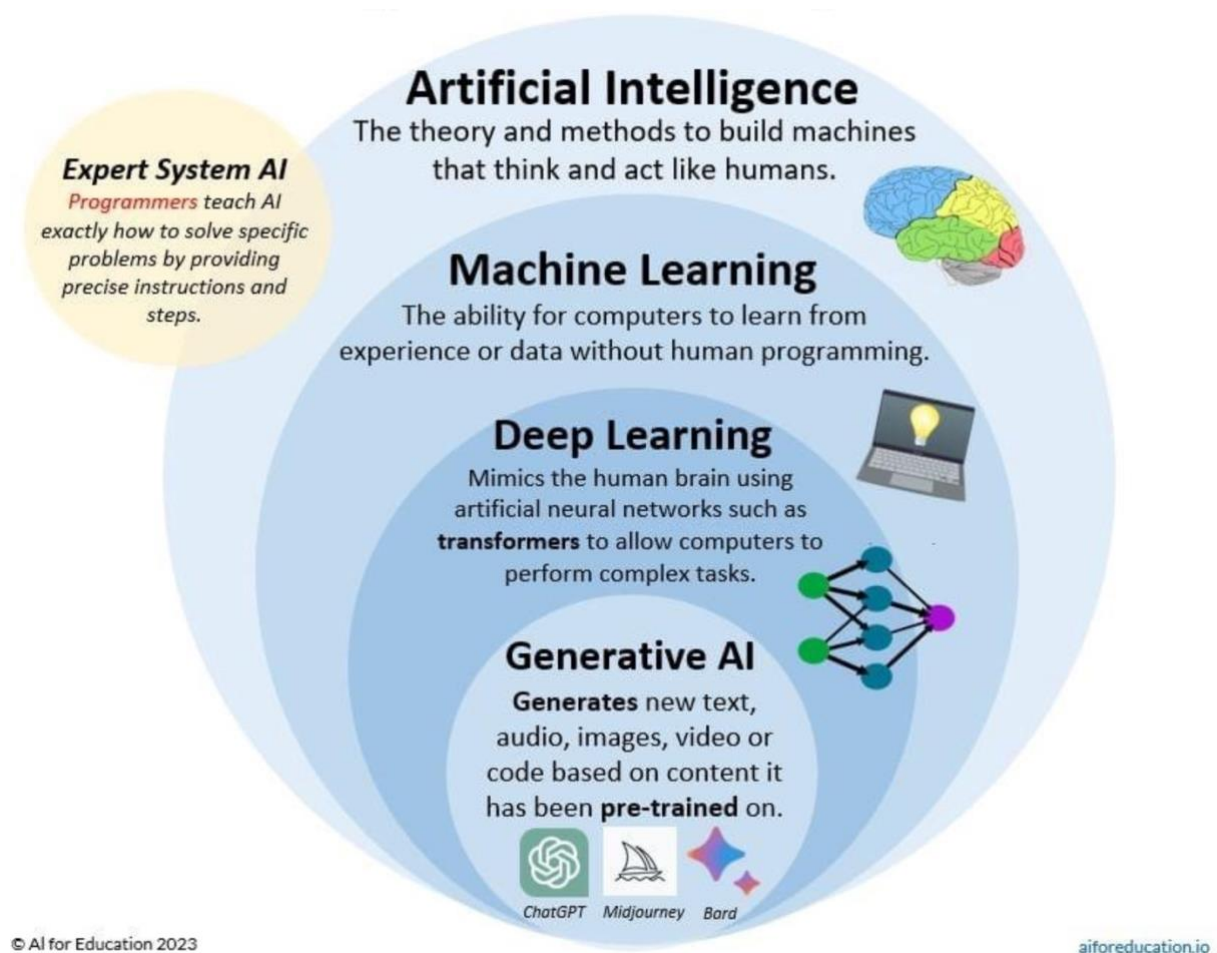


Figure 5.1 Overview of the animation of “Generative AI in a nutshell” by Henrik Kniberg (2024).

We usually distinguish between three main categories of AI: symbolic AI, machine learning, and generative AI. Symbolic AI, also called ‘GOFAI’ (Good Old-Fashioned AI or weak AI), consists of explicit rules and logic. An example is a chess program that calculates all possible moves through a tree structure and then makes the most strategic choice. Symbolic AI simply follows preprogrammed rules or learns patterns based on data. This approach was dominant until the late 1980s (Crevier, 1993). There is also a distinction here between so-called weak AI and strong AI. Strong AI is a hypothetical concept in which an AI system would be capable of human-like consciousness and general intelligence. It is also

called artificial general intelligence (AGI). Such a system would not only be able to perform tasks, but also reason, understand, learn, and even experience emotions like a human. For now, this is still the stuff of the future, if it ever becomes reality.

Viewing tip 5-1: would you like to see an example of how a strictly programmed system takes the rules too literally? Then watch the movie [Avengers: Age of Ultron](#) (2015), in which the AI Ultron runs amok because it interprets the assignment to protect humanity too rigidly.



p. 5-47

Figure 5.2 Depiction of the field of artificial intelligence and generative AI according to AI for Education (AI for Education, 2023)

With the rise of machine learning, a subcategory within AI, the approach changed radically. Instead of explicitly programming how a system should make decisions, models learn to do this themselves based on data. This made it possible to recognize patterns in data without human intervention at every step. Machine learning uses statistical techniques such as linear regression and decision trees (T. M. Mitchell, 1997).

A step further is deep learning, which uses artificial neural networks with multiple layers. These networks can learn complex patterns that were previously inaccessible to simpler algorithms. Think of recognizing objects in images or predicting words in a sentence. Deep learning forms the backbone of the most advanced AI applications we use today (LeCun et al., 2015).

Box 5.2 - Algorithms

In popular scientific and public discussion, all kinds of words are used interchangeably when it comes to AI, which creates confusion. For example, the word: algorithm. This term is used in different ways, and that is confusing because, in principle, an algorithm is nothing more than a step-by-step procedure. Even adding two numbers happens according to an algorithm. But training a language model also happens according to an algorithm. Such an algorithm is simply very complex and uses a great deal of data. You can then use a language model for a new application that again has its own algorithm, for example a chat environment, which we will discuss later in chapter 6.

Consider the statement: “the algorithm causes users of a social media platform to end up in an echo chamber.” This does not necessarily say anything about the AI approach used in the system. But it mainly says something about the way *the designers of social media platforms use the data in the system* to keep serving the user new posts.

Generative AI is a specialized form of deep learning. Instead of only classifying or predicting, generative AI creates new output: text, images, audio, or video. Systems such as ChatGPT, DALL·E, and Google Gemini are models trained on enormous amounts of data and can use that knowledge to generate original content. Generative AI is therefore not only reactive, but also creative: it writes poems, programs software, creates artwork, and simulates conversations (Bender et al., 2021). But generative AI only does this based on your creative input and the creative training data on which the model was trained. The output can therefore be creative, but it is rarely original if you do not give it your own creative twist.

p. 5-48

Box 5.3 - AI, statistics, and Big Data: differences and similarities

Big Data is sometimes confused or equated with AI and generative AI as well. That can be confusing. Big Data refers to the techniques and infrastructures used to store enormous amounts of data and process them quickly. That in itself has nothing to do with AI. It is, however, a prerequisite for developing and using AI.

In addition, some people wonder what the difference is between statistics and AI. In short: statistics is generally used to draw conclusions or test hypotheses based on structured data. AI, on the other hand, is strong at recognizing patterns in unstructured data, such as recognizing disease patterns on medical scans.

Generative AI is good at predicting the most likely answer to a prompt, based on a statistical model.

5.2 Neural networks and training

After exploring the AI landscape, we now look at the engine of generative AI: neural networks. Everything you see when you have an AI model write a text, generate an image, or make a summary is the result of what such a network has learned. But how does AI actually learn? And what happens behind the scenes when you enter a prompt? In this chapter, you will discover step by step how an AI model is trained and how a neural network functions.

5.2.1 How does a neural network work?

An artificial neural network is inspired by the human brain. It consists of layers of 'neurons' that process information. Each neuron takes input values, performs a mathematical calculation, and sends the result to the next layer. The connections between neurons have 'weights' that determine how important a given signal is. By adjusting these weights, the network learns to make increasingly better predictions.

In deep learning - the technique on which generative AI is based - these networks consist of dozens or hundreds of layers. Each layer recognizes a slightly more abstract pattern than the previous one. Where the first layer in an image, for example, recognizes only edges or areas of color, a higher layer can distinguish an eye, face, or chair (LeCun et al., 2015). In language models it works in a similar way: the first layers recognize individual words, the following layers understand grammatical structures or even the tone and intent of a text.

An important component of generative AI is the so-called transformer, introduced by Vaswani et al. (2017). This technique itself determines which words in a sentence are important for understanding the meaning correctly (see section 5.2.3). This allows models such as GPT-4 to analyze longer texts and give coherent answers. The T in GPT therefore stands for transformer.

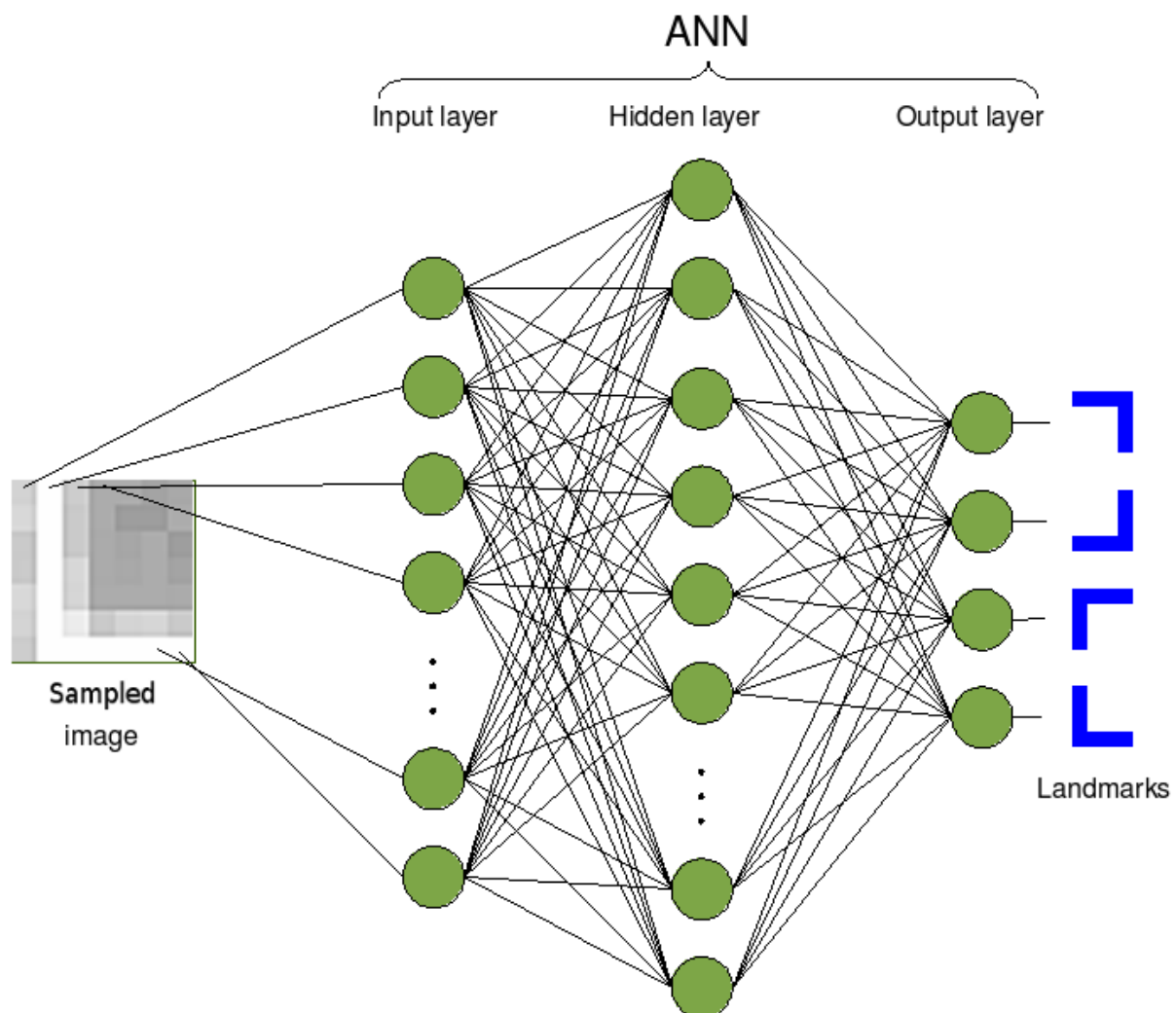


Figure 5.3 Visualization of how a neural network works to detect characteristic parts in a figure (source: Wikimedia/Cyberbotics Ltd.).

Box 5.4 - An audiovisual explanation of neural networks and training

How neural networks work is quite complex. Would you like to explore this further? Then watch the YouTube series [Neural Networks](#) by 3Blue1Brown (Sanderson, 2017).

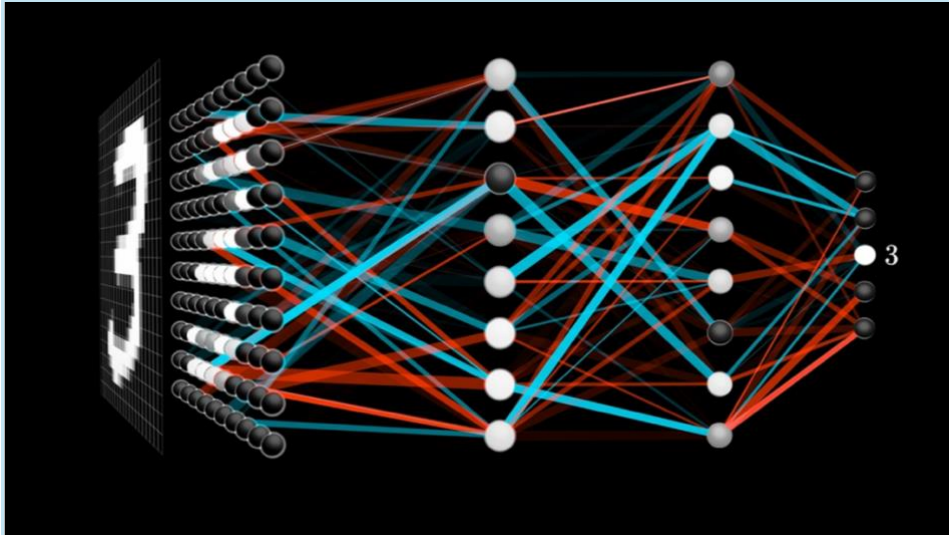


Figure 5.4 Image from the 3Blue1Brown collection by Grant Sanderson on neural networks (Sanderson, 2017).

Do you really want to understand it in depth? Then watch this lecture series on YouTube from Stanford University.

Lecture 1
Transformers
End-to-end walkthrough
CME 295

Stanford CME295: Transformers and Large Language Models
door Stanford Online
Playlist · 9 videos · 78.970 weergaven
This course explores the world of Transformers and Large Language Models (LLMs). You will learn the end-to-end training process.

▶ Alles afspelen

1 **Lecture 1 Transformers**
End-to-end walkthrough
CME 295
1:41:59

2 **Lecture 2 Transformer tricks**
CME 295
1:47:20

3 **Lecture 3 Large Language Models**
CME 295
1:48:45

4 **Lecture 4 LLM training**
Preparing LLMs
CME 295
1:47:27

5 **Lecture 5 LLM tuning**
RLHF with LLMs
CME 295
1:47:42

6 **Lecture 6 LLM reasoning**
RL reasoning with LLMs
CME 295
1:47:30

7 **Lecture 7 Agentic LLMs**
RLHF for agentic LLMs
CME 295
1:49:23

8 **Lecture 8 LLM evaluation**
LLM eval in a single benchmark
CME 295
1:49:25

9 **Lecture 9 LLM overview**
LLM overview, training recap
CME 295
1:51:31

Stanford CME295 Transformers & LLMs | Autumn 2025 | Lecture 1 - Transformer
Stanford Online · 354K weergaven · 2 maanden geleden

Stanford CME295 Transformers & LLMs | Autumn 2025 | Lecture 2 - Transformer-Based Models & Tricks
Stanford Online · 83K weergaven · 2 maanden geleden

Stanford CME295 Transformers & LLMs | Autumn 2025 | Lecture 3 - Transformers & Large Language Models
Stanford Online · 44K weergaven · 2 maanden geleden

Stanford CME295 Transformers & LLMs | Autumn 2025 | Lecture 4 - LLM Training
Stanford Online · 42K weergaven · 2 maanden geleden

Stanford CME295 Transformers & LLMs | Autumn 2025 | Lecture 5 - LLM tuning
Stanford Online · 19K weergaven · 1 maand geleden

Stanford CME295 Transformers & LLMs | Autumn 2025 | Lecture 6 - LLM Reasoning
Stanford Online · 25K weergaven · 1 maand geleden

Stanford CME295 Transformers & LLMs | Autumn 2025 | Lecture 7 - Agentic LLMs
Stanford Online · 41K weergaven · 1 maand geleden

Stanford CME295 Transformers & LLMs | Autumn 2025 | Lecture 8 - LLM Evaluation
Stanford Online · 25K weergaven · 1 maand geleden

Stanford CME295 Transformers & LLMs | Autumn 2025 | Lecture 9 - Recap & Current Trends
Stanford Online · 56K weergaven · 1 maand geleden

5.2.2 What is machine learning?

The next step is to understand how neural networks actually learn: we call this process machine learning. Machine learning is the process by which computer systems learn from examples. Instead of programming explicit instructions, we give the model a large amount of data as training data - for example millions of texts - and let the system discover for itself what the patterns are. That training data is the foundation on which the model learns, for example, what is grammatically correct, which words often occur together, and what a particular writing style sounds like (M. Mitchell, 2019).

This learning process is based on trial and error. The model makes a prediction (for example: what is the next word in a sentence?), is told whether the answer was correct, and adjusts itself. This error correction happens using so-called 'backpropagation': the computational model sends an error signal back through the network and adjusts the internal weights on that basis (Goodfellow et al., 2016).

5.2.3 How does a language model learn by guessing the next word?

To understand how a language model learns, you can compare it to an advanced prediction machine that learns from millions of examples. During training, the model is shown sentences in which it must repeatedly predict what the next word is, complete with probabilities.

Step 1: The training sentence

Suppose the model is given the sentence: "The cat sits on the mat"

Step 2: The prediction game with probabilities

At first, the model sees only: "The". It predicts with probabilities:

- "dog" (15% chance)
- "cat" (12% chance) ✓ correct!
- "man" (10% chance)
- "car" (8% chance)
- ... and thousands of other words with lower probabilities

Why is "dog" wrong? Not because dogs cannot exist after "The", but because in this specific training sentence the correct answer was "cat". The model now learns that it should increase the probability of "cat" after "The" slightly.

Step 3: The attention mechanism - what should I pay attention to?

With longer sentences it becomes harder. Suppose the model sees: "The black cat with the white spots sits on the"

The attention mechanism helps the model 'remember' that:

- "cat" is the subject (not "spots")
- "sits" is the action
- We are probably looking for a place where the cat is sitting

This mechanism assigns weights to different words. The word "cat" gets a lot of attention (0.8), while "with" gets little attention (0.1).

Step 4: From training to use - the role of temperature

After training, the model can generate text. For the sentence "The cat sits on the" it has these predictions:

- "mat" (45% chance)
- "couch" (20% chance)
- "chair" (15% chance)
- "table" (10% chance)
- "moon" (0.1% chance)

This is where **temperature** comes into play:

- **Low temperature (0.2):** the model almost always chooses "mat" - safe but boring.
- **Normal temperature (0.7):** usually "mat", sometimes "couch" or "chair" - natural variation.
- **High temperature (1.5):** it may even choose "moon" - creative but possibly nonsensical.

p. 5-52

Step 5: Learning from billions of examples

After seeing billions of sentences, the model learns complex patterns:

- "The prime minister announced that..." → policy will probably follow (70%)
- "The cat sits on the..." → probably a piece of furniture or a mat (85%)
- "The research shows that..." → probably a scientific conclusion (75%)

The result

The model becomes a master at predicting likely word combinations. The attention mechanism helps it 'remember' relevant context over longer distances in the text. By adjusting the temperature, we can determine how 'safe' or 'creative' the output becomes.

But remember: the model does not really 'understand' that cats are animals. It only knows that in texts where "cat" appears, words such as "meows" (65% chance), "mice" (45% chance), or "tail" (40% chance) also often appear. This explains why AI can sometimes sound convincing and still say factually incorrect things - it is based on statistical patterns, not real understanding.

5.2.4 Different ways AI is trained

In supervised learning, examples with labels are used to learn to recognize patterns (Hastie et al., 2009). Imagine that you want to teach a computer to recognize whether an email message is spam or not. You then give the system a large collection of emails, and with each email you provide a clear label: this is spam, and this is not spam. This label helps the algorithm establish relationships between characteristics of the email (such as words, sender, links) and the label.

Who does the work?

A data scientist or machine learning engineer sets up the model and collects and labels the data. Sometimes the labeling is done manually, for example by human data labelers. If you need thousands or millions of examples, this is often outsourced to specialized companies. In chapter 9 we discuss the social aspects of this.

How much time does it take?

Building a well-functioning model can take weeks to months, depending on the size of the dataset, the complexity of the problem, and how well the data is prepared. Just collecting and labeling data can already take many tens to hundreds of hours. For creating LLMs or vision models that can recognize images, this can run into thousands or even millions of hours.

5.2.5 Unsupervised Learning - learning without examples

Unsupervised learning enables systems to discover patterns without predefined labels (Bishop & Nasrabadi, 2006). Imagine this: you give an algorithm customer data (such as age, purchase history, and location), but you do not say which customers are loyal customers or which are high-risk customers. The model must then try to recognize patterns by itself. For example: this group of people comes back often, and that group only once. Or: this group of people returns products more often, and this group does not.

Who does the work?

Here too, data scientists build the model and data labelers work on the training. The work lies mainly in collecting and cleaning the data, choosing suitable algorithms (such as clustering or dimensionality reduction), and interpreting the patterns.

p. 5-53

How much time does it take?

Unsupervised learning can be faster than supervised learning because you do not need labels. But interpretation by data scientists takes more time, because the results are not self-evident. Count on a few days to a few weeks for an average analysis. For LLMs or general vision models (models that can recognize and generate images without prior knowledge), this increases enormously.

5.2.6 Reinforcement Learning - learning through reward and punishment

In reinforcement learning, a system learns optimal behavior through interaction with an environment and feedback via rewards (Sutton & Barto, 1998). This way of learning is a bit like how people or animals learn. You have the computer perform a task (for example making a robot walk or play a game) and give a reward when it goes well. If it goes wrong, it gets no reward or even 'punishment.' Based on that feedback, the system learns what works and what does not.

Who does the work?

Reinforcement learning is usually carried out by specialized AI researchers or engineers because it is technically complex. It often requires simulations, a great deal of computing power (Graphics Processing Units - GPUs), and experience with frameworks such as OpenAI Gym, Unity ML, or TensorFlow Agents.

How much time does it take?

This is often the most time-consuming. Training a good reinforcement learning model can take days to weeks - sometimes longer - depending on the complexity of the task. Think of hundreds to thousands of hours of compute time and preparation, including building the simulation environment.

Learning method	With or without labels	Who carries it out?	Average time investment
Supervised Learning	With labels	Data scientists + data labeling teams	Several weeks to months
Unsupervised Learning	Without labels	Data scientists	Several days to weeks

Reinforcement Learning	Feedback-based	AI researchers / engineers	Weeks to months (enormous compute time)
-------------------------------	----------------	----------------------------	---

Table 5.1 Overview of machine learning methods.

Box 5.5 - What does the work of a 'data labeler' look like?

In De Groene Amsterdammer, Jeroen van Bergeijk writes (2025) about the work of a data labeler from personal experience. An AI data labeler, also called an AI trainer or AI annotator, works as a freelancer for companies such as Outlier to improve language models such as ChatGPT. The work consists of devising prompts to which the AI model gives two answers, after which the data labeler must assess which answer is best and why. The assessment is made using seven different dimensions such as writing quality, truthfulness, and harmfulness, with each violation substantiated. The goal is to make the AI model make mistakes by creating challenging prompts with many conditions and constraints; this is how models are improved and unwanted output is prevented.

There is often a great deal of secrecy surrounding this work - data labelers often do not know which specific AI model they are working for and must sign extensive nondisclosure agreements. Pay ranges from 20 to 30 dollars per hour in the Netherlands, but in low-wage countries it can be between 1 and 5 dollars per hour (Perrigo, 2023). The availability of work is highly unpredictable - sometimes work is available for days on end, and then again nothing at all for weeks. Many data labelers become frustrated by the time pressure, the lack of feedback, and the unclear quality requirements. More about this in chapter 9

p. 5-54

The learning process of a language model proceeds *technically speaking* in roughly six steps:

1. **Tokenization.** The input (text) is divided into smaller units: tokens. These can be words, but also syllables or even individual characters. For example, the sentence "The sun is shining" is divided into: ["The", " sun", " is shining"]. Division can be done in multiple ways, but a rule of thumb is that 1 token is about 3-4 letters long (CriticalMynd, 2025). This means that, for example, a blog of about 750 words is converted into about 1,000 tokens, an essay of 3,000 words into about 4,000 tokens, and a book of about 750,000 words into 1 million tokens.
2. **Vectorization.** Each token is converted into a sequence of numbers: a vector. This vector contains information about the position of the word in a language space - comparable to coordinates on a map.
3. **Embedding.** All vectors are combined in a so-called embedding space that represents the meaningful relationships between words. Words with similar meanings are located close together in this space.
4. Training via **backpropagation.** The network learns by making predictions and correcting errors. If, for example, it predicts the word "sun" after "The", and the correct answer was "moon", then the model adjusts the internal parameters to make better predictions in the future.
5. **Loss functions** measure how good or bad a prediction is. They steer the learning process by saying: "You were off by 30%."
6. **Fine-tuning** is an additional training phase in which an already trained model is adapted to specific tasks or contexts - for example, legal or medical applications. In this process, the weights in the model are adjusted and a copy of the language model is created.

How it works

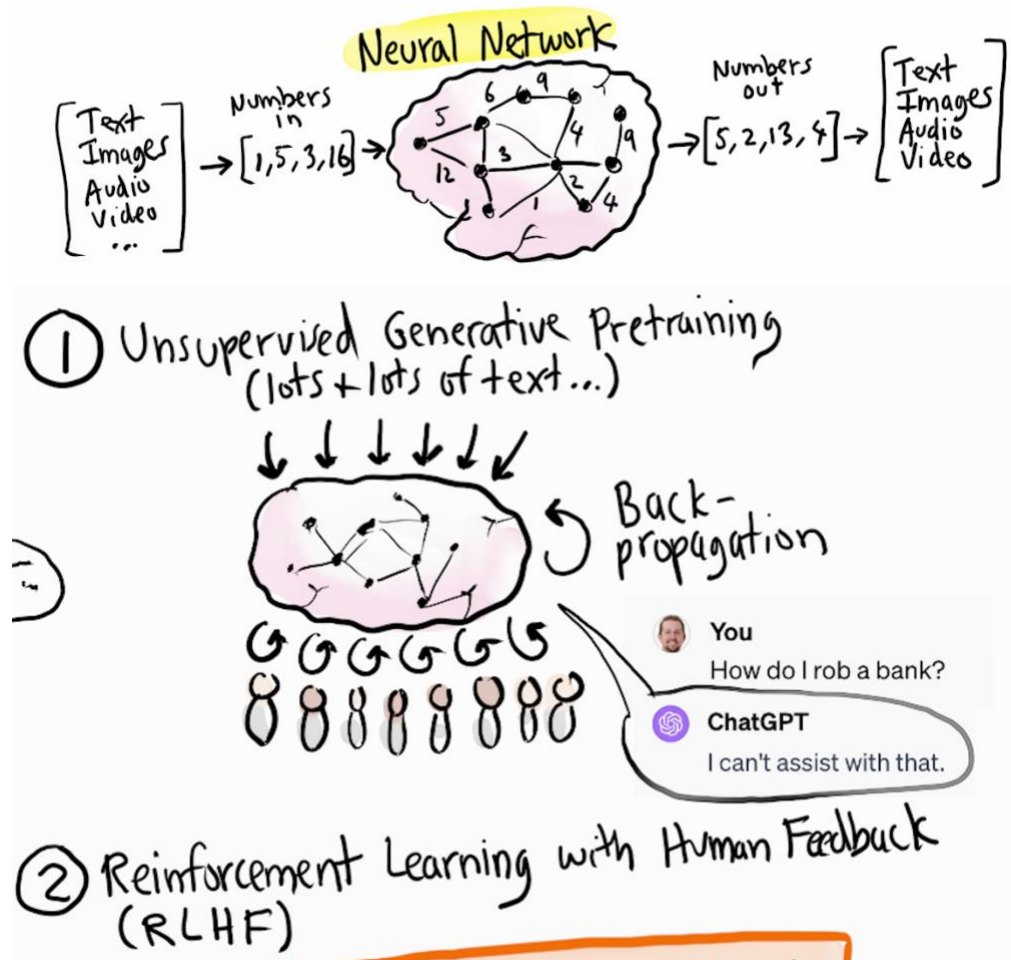


Figure 5.5 Visualizations of neural networks and training based on the video “Generative AI in a nutshell” by Henrik Kniberg (2024).

5.2.7 Diffusion Models (Diffusers) – generative AI for images

A diffusion model is used specifically for generating images. Diffusion models, as first described by Ho et al. (2020), are a type of generative AI that is trained by gradually adding noise to an image or audio file. Think of this noise as static interference: elements that do not belong in the original image, such as distortion or random colors. This added noise is called ‘Gaussian noise’ because it follows a normal distribution - a perfect bell-shaped curve.

In the training phase, more and more noise is added until the file is completely unrecognizable. At the end, the image or sound consists only of noise. This phase is called diffusion. What the model still knows about the original image is stored only in its memory. Then the reverse process begins: the model tries step by step to remove the noise again and reconstruct the original image as well as possible. This reconstruction phase is called reverse diffusion. In doing so, the model uses what it has previously learned about patterns and structures in other images. Through this process, the model learns to generate new images that resemble the structures it knows from the training data.

During this learning process, the model becomes better and better at predicting which noise has been added. By analyzing the difference between the original image and the distorted version, the model

learns which patterns are noise and which belong to the image. This continuous interplay between adding and removing noise forms the core of how diffusion models function and learn. In this process, AI is taught by data labelers what is and is not correct: for example, pictures in which a dog would get 5 legs during diffusion instead of 4 legs.

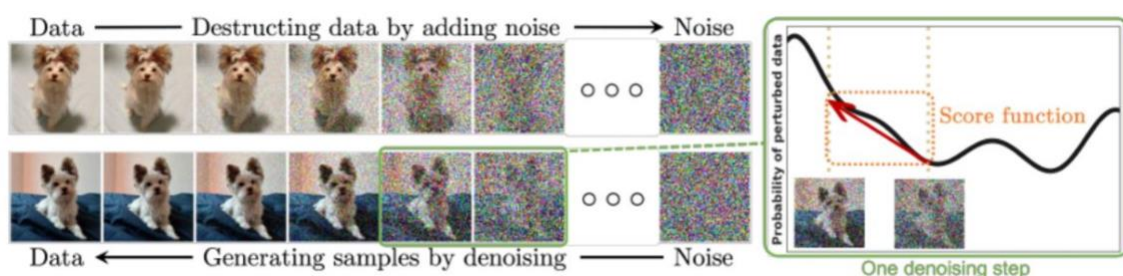


Figure 5.6 Images used to show the diffusion process for generating and recognizing images from Yang et al. (2024).

5.3 Bias and data quality

p. 5-56

One of the most discussed and at the same time often misunderstood aspects of AI is the risk of bias: systematic distortions in the output of an AI system that lead to inequality, misrepresentation, or even discrimination. At first glance, AI seems neutral: it is 'just an algorithm.' But in reality, AI systems are only as good (or as bad) as the data on which they were trained. In this section, you will discover how bias arises, why it is so persistent, and what you can do to recognize and reduce it.

5.3.1 Origins of bias in AI

Bias in AI usually does not arise from ill intent on the part of developers, but because datasets are used in model training that are unevenly distributed in terms of content. Think, for example, of a language model trained mainly on English-language texts from American websites: that model will perform especially well in American contexts and may have trouble with idioms or cultural references from other parts of the world (Bender et al., 2021).

A classic example comes from the medical world: an AI model that had to recognize skin conditions performed poorly on people with darker skin tones, because the training data mainly consisted of photos of white skin (Buolamwini & Gebru, 2018). So the problem was not in the model itself, but in the representation of the data on which it was trained. If the training data mainly includes men, or people from a certain income group, or texts from a certain ideology, then the model will reproduce those imbalances in the output. Suppose you are conducting sociological research into preferences in pop music; your results via AI are likely to be colored. This is because the data on which the AI is trained already contains a certain bias, for example mostly Western data.

Box 5.6 - Data is never neutral

People often assume that data is objective. But data is always a reflection of choices: what is measured, who is included, which categories are used, and how is something labeled? A chatbot trained on online forums (Reddit, for example, is a widely used source for AI training) may use less inclusive language than a chatbot trained on academic texts. The labels assigned to data - for example in supervised learning - are also subjective: they arise from human interpretations.

AI models are therefore not only technical systems, but also a reflection of the world in which they came into being, with all the associated culture, politics, prejudices, and discrimination (Crawford, 2021b).

5.3.2 Strategies for reducing bias

Although you can never completely rule out bias, there are ways to limit its impact, such as:

- **Dataset diversification:** collect training data from multiple sources, with attention to representing diversity in gender, ethnicity, language, region, age, and socioeconomic background.
- **Bias audits:** conduct systematic tests on AI output to determine whether certain groups are systematically disadvantaged.
- **Transparency:** document how the model was trained, which data was used, and which choices were made in the process (Model Cards, datasheets).
- **Human oversight:** always have sensitive decisions - such as in selection procedures or health care - checked by people.

5.3.3 Representativeness and accountability

An important question is: who is the model intended for, and who remains out of view? AI is often developed in Silicon Valley, but used all over the world. Without explicit attention to context, this can lead to applications that do not respect cultural differences, or even have harmful consequences. Think of automatically generated images that reinforce stereotypes, or summaries that leave out certain perspectives.

Responsible use of AI therefore means not only technical precision, but also social reflection. Who determines what counts as 'normal' in a dataset? Who has access to the technology, and who does not? And who bears the consequences of mistakes?

Box 5.7 - Where does machine learning go wrong?

Supervised learning: examples with a label are used here, but labels are not always objective. What is neutral can differ by culture or context.

Unsupervised learning: there are no labels here, but the model groups data based on statistical similarities. This can reproduce existing inequalities. Take the example of a system designed to recognize fractures in the ankle joint. The system turned out to work quite well, but not because of the fractures in the images, but because of its recognition of wrinkled skin in older people. Older people are more likely to have fractures.

Reinforcement learning: when human feedback is used, there is a risk that the human preferences contain bias, and that this becomes embedded in the system.

Bias can arise at any stage: when collecting data, selecting features, training the model, or even interpreting the output.

5.4 AI hallucinations

Suppose you ask a language model who the president of France is. You expect Emmanuel Macron as the answer. But the model replies: François Hollande. Or even worse: Victor Hugo. These kinds of incorrect or made-up answers are known as AI hallucinations. The word 'hallucination' is used here because the AI appears to invent factual information based on what sounds plausible, without that information actually being true. In this section, you will discover what AI hallucinations are, why they occur, how to recognize them, and what implications they have for the use of generative AI.

5.4.1 What are AI hallucinations?

An AI hallucination is output that seems grammatically and stylistically correct, but is factually incorrect, misleading, or completely made up (Maynez et al., 2020). In language models such as ChatGPT, this means that the model generates sentences that do not correspond to reality, but do sound as if they could be true. For example, a chatbot may cite a non-existent scientific study or summarize an article that does not exist. In image generators such as DALL-E, hallucinations can lead to illogical images, for example a hand with extra fingers or non-existent objects.

Hallucinations are one of the biggest challenges to the reliability of generative AI, especially in contexts where factual accuracy is essential, such as scientific writing, medical advice, or legal support.

5.4.2 Why do hallucinations arise?

The fundamental mechanism of a large language model is based on probability calculation. The model repeatedly predicts which token - a word or part of one - is likely to follow based on all preceding tokens. This process, also called 'next-token prediction,' is purely statistical. The model does not know what is true or what exists; it only knows what sounds likely within the context (Brown et al., 2020).

If the training data is insufficient on a particular subject, or if the context of the prompt is vague, the model can make incorrect assumptions. Sometimes it invents sources because it has learned that an article with a year and an author's name is a good format, without actually checking whether that article exists. The model's so-called temperature setting - which determines how creative the model is allowed to be - also plays a role here: the higher the temperature, the greater the chance of surprising but unreliable output.

5.4.3 Consequences of hallucinations

In academic or professional contexts, AI hallucinations can have serious consequences. A student who cites a non-existent source risks being accused of fraud. A doctor who relies on AI for a rare diagnosis may choose the wrong treatment. Even in creative contexts, such as storytelling or design, hallucination can be problematic if the user expects a certain level of consistency.

That is why human intervention is always necessary: the user must check the information, verify the sources, and remain aware of the system's limits. Chatbots are not oracles, but advanced guessing machines. Without a critical eye, their output can mislead, despite a professional or convincing tone.

Box 5.8 - Techniques to combat hallucinations

Suppliers of LLMs recognize the problem of hallucinating and want to counteract it. For example, through fine-tuning. In doing so, you take an existing language model but adjust the parameter weights through specific training.

Another method works on the basis of so-called Retrieval Augmented Generation (RAG) (Lewis et al., 2020). In this approach, documents are fed to the system (for example PDFs, website texts, text files), with the LLM tailoring the response to the content of these documents. This allows you to add a lot of basic information to the bot as it's being built.

With RAG, a search system automatically selects the most relevant passages from the documents and includes them in the prompt, which is then sent to the LLM. With prompts, you determine the prompt text entirely on your own. Because of the

automatic search, an RAG approach does carry a higher risk of hallucination than direct prompting.

In addition, more and more techniques are being applied in which models are forced to start an investigative conversation with themselves for better outcomes (so-called reasoning).

5.4.4 Recognizing and dealing with hallucinations

There are several strategies for recognizing and dealing with AI hallucinations:

- Check sources: ask the model for the exact URL or DOI of an article. Invented sources often have no working link. A source can still be real even if the link does not work, however. Therefore, especially ask for the title, author, year, and publication, and look them up yourself.
- Ask for justification: have the AI explain the reasoning behind a particular conclusion. Vague or circular argumentation is a warning signal.
- Use fact-checking tools: combine AI output with tools such as Google Scholar, Consensus, PubMed, or Wikipedia to verify claims.
- Lower the temperature: if you have control over the model settings, lower the temperature to make the output more predictable and factual.

p. 5-59

Box 5.9 - Why AI believes it is right

Hallucinations are not a flaw of a defective model, but a logical consequence of how language models are designed. They have no world knowledge or memory of real events. They do not think like humans. What they generate is based on patterns, not facts. Even though the sentence "Victor Hugo was president of France in 2018" sounds grammatically perfect, it is pure fiction - statistically derived, not truthful.

In addition, it is important to realize that LLMs themselves do not have access to the current state of affairs, but chat interfaces such as ChatGPT often do because they can search the internet or specific databases. Through techniques such as RAG and by also making connections via the chat interface with search engines or, for example, specific databases.

5.5 The 'Black Box' of AI

Although generative AI often seems impressive because of its fluent language and quick responses, many of the underlying processes remain invisible and difficult to understand - even for the developers. This phenomenon is known as the 'black box' problem. An AI model may arrive at accurate predictions or generate convincing texts, but it is often not clear how or why it does so. In this section, we dive deeper into the nature of this opacity, the attempts to clarify it, and its ethical and social implications.

5.5.1 Why is AI so opaque?

The complexity of modern neural networks is hard to overstate. Models such as GPT-4 contain hundreds of billions of parameters - weights and connections between neurons that together determine the model's behavior (OpenAI, 2021). When you enter a prompt, thousands of calculations are carried out in fractions of a second. Each of these calculations contributes a little bit to the final result, but no single individual step is decisive on its own. This makes it virtually impossible to reconstruct exactly which paths are responsible for the generated output.

In addition, these models learn via ‘backpropagation’ based on enormous amounts of data. They internalize patterns without following explicit logic or rules. The result is that they function like a kind of complex intuition: extremely efficient, but difficult to explain. This lack of transparency stands in the way of control, trust, and accountability - especially in contexts with high social risks, such as diagnosing a medical condition, making decisions about granting benefits, or admission to educational programs.

5.5.2 Attempts at explainability

AI may sometimes seem like a mystery to you: an answer comes out, but you do not know how it was produced. To counter that 'black box' feeling, the research field of ‘explainable AI’ (XAI) emerged. Researchers are looking for ways to make it clearer how AI arrives at a decision. They do this, for example, by showing what the model pays attention to in a text, or by temporarily replacing a complicated model with a simpler model that is easier to understand (Lipton, 2018).

Two of the techniques for this are, for example:

- SHAP (SHapley Additive exPlanations), which tries to determine which input features contributed the most to a particular output.
- LIME (Local Interpretable Model-agnostic Explanations), which locally approximates the behavior of a complex model with a simpler, interpretable model.

p. 5-60

You may forget these terms again after this, but what is important to remember is that these kinds of methods provide only a limited view. They give an impression of what the model may have done, but no conclusive proof. In language models such as GPT, this is even more difficult, because the input is constantly changing and the outcome depends on long contexts and probability calculations.

5.5.3 Ethical implications of opacity

The black-box nature of AI raises fundamental questions about responsibility and liability. Who is responsible if an AI system makes a mistake - for example, a discriminatory decision in a hiring procedure or an incorrect medical recommendation? If we cannot explain why a model does something, can we really use it in decisions that affect people’s lives?

Cynthia Rudin (2019) argues that black box models should not be used in sensitive domains such as justice, medicine, or credit provision. Instead, she advocates interpretable models that may be slightly less powerful but whose operation is understandable and testable.

Moreover, opacity leads to distrust. Citizens, policymakers, and professionals want to understand why a system arrives at certain conclusions. Without explainability, AI systems lose legitimacy. Transparency and explainability are therefore not luxuries, but conditions for socially responsible AI use.

5.6 Self-study questions

5.6.1 Check questions

1. What is the difference between supervised learning, unsupervised learning, and reinforcement learning? Give a brief example of each.
2. What are AI hallucinations and why do they arise? Name two strategies for recognizing hallucinations.
3. Explain what the 'black box' problem of AI is and why this raises ethical questions.

5.6.2 Reflection questions

4. Chapter 4 describes how bias in AI systems arises from unevenly distributed datasets. Think of a concrete example from your own field of study where this could be problematic. How would you deal with this as a future professional?
5. This chapter compares using AI to driving a car without understanding how the engine works. To what extent do you think users of AI systems should understand the underlying technology? What are the advantages and disadvantages of more technical knowledge for ordinary users?
6. Because of AI use, some words suddenly get used much more often. For example, in English it is noticeable that the word 'delve' suddenly appears much more often in texts. Some people think this could be because certain LLMs are mainly trained by people from certain regions. [For instance, the word 'delve' might be more common in Kenya, where many LLMs are trained.](#) What do you think? Argue your answer.

5.6.3 Suggested answers

1. Supervised learning uses labeled examples to learn patterns (for example labeling emails as spam or not spam). Unsupervised learning looks for patterns without labels (for example grouping customers based on purchasing behavior). Reinforcement learning learns through reward and punishment (for example teaching a robot to walk through positive feedback for good steps).
2. AI hallucinations are convincing-sounding but factually incorrect answers generated by AI. They arise because AI statistically predicts what is likely to come next, without real knowledge of facts. You can recognize them by checking sources, asking for justification, using lower temperature settings, and using fact-checking tools.
3. The black box problem means that we cannot explain how AI arrives at specific decisions, even when the results are accurate. This raises ethical questions about responsibility, transparency, and liability, especially in sensitive domains such as health care or the justice system, where explainability is crucial.
4. Personal reflection desired. Think, for example, of medical diagnoses that may contain prejudices about certain patient groups, or HR systems that discriminate in job applications. As a professional, you can be aware of these risks, require diverse datasets, conduct regular audits, and always maintain human oversight in important decisions.
5. Personal opinion desired. Arguments for more knowledge: better critical use, more trust, better quality control. Arguments against: technical complexity is not accessible to everyone, and the focus should be on practical use. A middle ground could be a basic understanding of possibilities and limitations without in-depth technical details.

6 Chat software: a neat wrapper around your LLM

A language model by itself is not usable for end users. That is why the language model is used within another program so people can work with it: especially in chat programs, but also anywhere a generative AI feature is built into an existing program. For the end user, those programs often run in browsers to make them easy to use. But all the heavy computation takes place in data centers.



p. 6-62

In this chapter you will learn

- That, technically speaking, a language model is simply a very large file.
- That a chat environment is the wrapper around that file so you can use it easily.
- That continually resending earlier questions and answers creates the impression that you are having a conversation with a language model.
- That language models require a great deal of computing power and data storage, but that a (small) language model can also run on your own computer.
- That language models can also send messages to one another, to other systems, or to themselves to detect errors so they become 'smarter' (Agentic AI)

Takeaways

- You are aware that vendors design chat software to make you, as a user, want to engage with it, for example by always being very friendly and understanding (sycophancy).

Keep in mind: be aware that a provider of chat software can access all the data from a conversation, including the files you send to it (upload), and can use that data at its own discretion. You need to read the provider's terms carefully to know what happens to your data.

What do you think?

Would it be better if chat software were less friendly and also did not present itself as a human being, but as a technical system? For example: the software is not ‘thinking,’ but ‘processing data.’

Imagine ...

You scroll through Copilot's answer to a research question you asked. You think to yourself: "Theoretically I understand how these systems work, something with neural networks, but what practically happens to my input? I type in questions, but where does that question actually end up?" And you remember your instructor's warnings: "Never share sensitive data through these systems." Could your information be stolen? Resold? And is that always the case? Time to take a look at how these systems work in practice and what that means for your use.

6.1 Infrastructure

In general, you communicate through a chat interface with AI systems that contain Large Language Models (LLMs). This technology seems simple—you type a question and get an answer—but behind the scenes it requires a complex technical infrastructure. Many people think they are talking directly to an LLM and that their data is processed only by the model, and that the model automatically learns from it. In reality, all communication passes through the chat interface, which is managed by another party. This leads to misunderstandings, for example about:

- who has access to your data;
- whether your data is used for model training;
- where your data is stored and for how long.

It is therefore essential to understand the difference between the LLM file (the model) and the provider of the chat interface (the intermediary layer, in other words: the wrapper, the packaging).

6.2 The LLM is a large file

An LLM is a single file that you could, in principle, put on your own computer. The file contains billions of parameters trained on large amounts of data. This file does not include built-in chat functionality: it is a file that, given a particular input (prompt), generates a single response based on the GPT method (Generative Pretrained Transformer, see Box 1.1 - What do we mean by 'GPT?'). The model itself remembers nothing from previous interactions; it is what is called stateless. The file also has no knowledge of, for example, the current date (and platforms offer additional techniques and functions that do allow it to 'know' the current date or other sources).

Box 6.1 - LLM Models

The number of LLM models you can use is growing quickly—so quickly that it is almost impossible to keep up. Each model has its strengths, weaknesses, and specialties. The best-known models are those from OpenAI (ChatGPT-4o, 5), Google (Gemini), Anthropic (Sonnet, Opus), DeepSeek (DeepSeek), Qwen (Alibaba Cloud), xAI (Grok), LLAMA (Meta), and Mistral (Mistral). These are the largest and most broadly applicable models, with the most parameters.

If you want an overview of available models, take a look at HuggingFace's website: <https://huggingface.co/models>. Or look in Table 6.1.

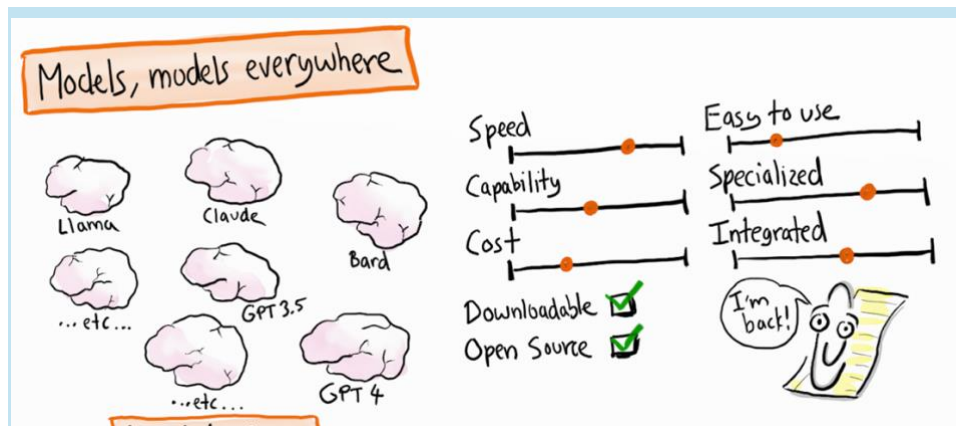


Figure 6.1 Overview of different LLM models and their characteristics from the video “Generative AI in a nutshell” by Henrik Kniberg (2024).

6.3 Chat interface and chat history

The experience of an ongoing conversation with a generative AI is created by an intermediary layer: the chat interface (for example, a web app or app—think of Copilot, ChatGPT, Gemini, or Claude).

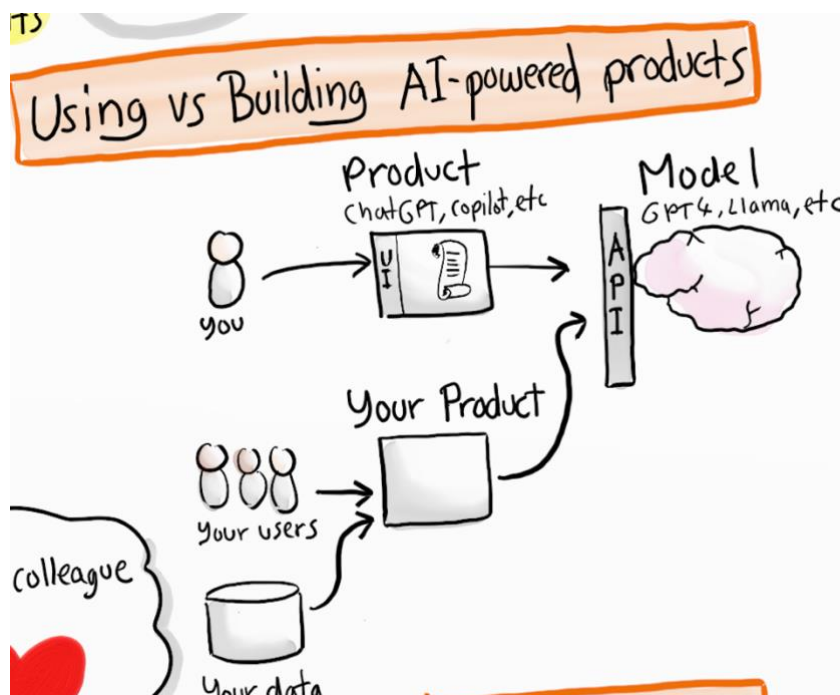


Figure .6.2 Graphic representation of how users interact with an LLM through an intermediate product. The user sees an interface (UI) in a product, and that product communicates with the model and back again via Application Programming Interface standards (API).

This chat interface stores the full chat history and, with each new question, sends the relevant chat history back to the LLM file. This is called conversational memory. This principle emerged alongside the rise of LLMs (Pinecone, 2025) and is essential for creating coherent conversations with AI.

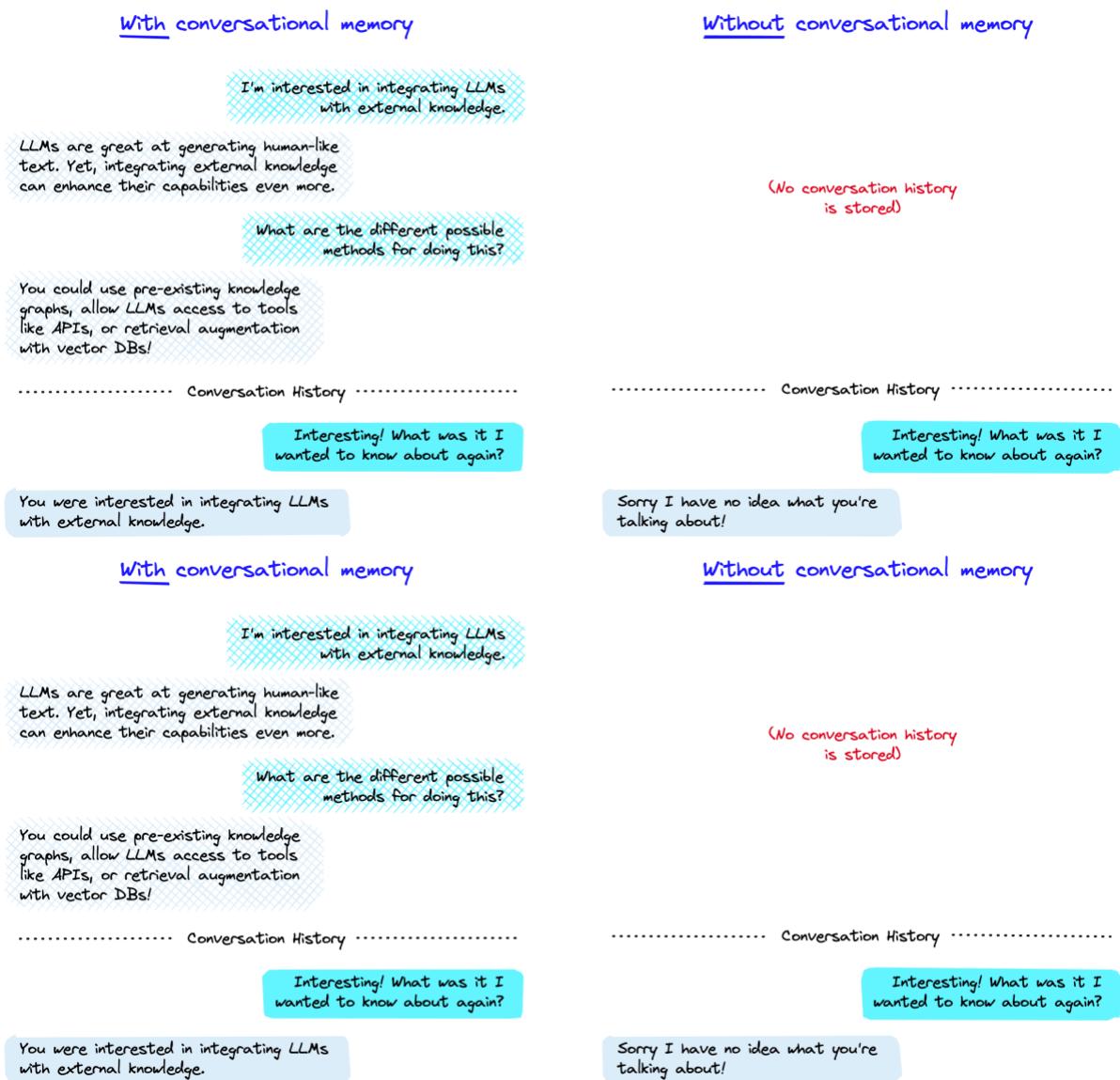


Figure 6.3 Example of conversational memory versus no conversational memory (Pinecone, 2025)

So the LLM keeps seeing the entire conversation (or the part that fits within the context window), which makes it seem as though the model has a memory and understands well what you want to ask. Without sending the first question and the answer back along with it, the LLM does not know what the conversation is about.

Box 6.2 - Context window: when does the chat history break down

When you keep talking to ChatGPT in the same chat box, that sometimes seems convenient because the chat box carries everything you enter forward into the rest of the conversation. But you may notice that over time ChatGPT forgets what you said at the beginning of the conversation. This is due to the so-called context window, the model's memory. GPT-3.5, for example, could remember only 2048 tokens, whereas GPT-4 can handle more than 32,000. This determines how large the text files are that a chat system can process and how far back a model can look in a chat history. The larger and more numerous the questions and responses, the

faster the context window fills up and the chat environment loses control or becomes less coherent.

6.4 Preconfigured chatbots and projects

Using general chat environments such as Copilot and Gemini, you can also create preconfigured bots. For example, you can create so-called Custom GPTs: a fixed prompt, optionally with an added knowledge source, that you can reuse again and again. This saves time because you do not have to re-enter those instructions every time, and it leads to more consistent responses from the language model.

In addition, in platforms such as Gemini , you can also create a project (Gemini calls them Gems). A project is a kind of workspace in which you bring together files, sources, and rules. Within that project, the AI responds specifically to that context each time, so you do not have to re-enter it every time in a new chat within the project. A bot mainly defines a role or way of working (such as “always answer in an academic style”), whereas a project is mainly about organizing your own material and agreements.

For example: if you want to query one document in different ways, you can turn it into a Gemini Gem and get started right away. As a researcher, you can specify in the instructions which domain the model should primarily draw knowledge from and that the answers should always include references. Or, in image generation, you can describe a fixed style so that every new image has the same look. In a project, you can also store your files and context on top of that, so you do not have to upload them again and again.

For clarification: Google calls this mechanism of preconfigured chatbots Gems, Anthropic calls it Projects, Copilot calls them Agents and EduGenAI calls this a Persona.

You also increasingly see such built-in, preconfigured bots in ordinary programs such as Word or Google Docs. Often you do not even notice it: you simply use a menu option to summarize text, translate it, or automatically create slides based on a web page.

Based on the technology, you can also imagine that you—or companies themselves—create and offer such an intermediary layer. Below you see a number of examples.

Provider	Product/intermediary layer	Models	Safe for VU Amsterdam?	Origin (Country, Continent)
SURF / Npuls	EduGenAI (likely available in January 2027)	Mix of open-source (such as Mistral, LLaMA) and commercial models (including GPT-4, Claude)	High	Netherlands, Europe
Microsoft	Copilot for Web/ Bing Chat	GPT-4, GPT-4o (via OpenAI), with web browsing and multimodality	High (via SURF license)	United States, North America
Microsoft	Copilot for Microsoft 365	GPT-4 (via OpenAI), integrated into	Not available for VU	United States, North America

		Word, Excel, Outlook, Teams		
Google DeepMind	Gemini	Gemini 1.5 Pro, Gemini Flash (multimodal, real-time web access)	High (Google Workspace for Education)	United States, North America
Google DeepMind	Gemini Notebook	Gemini 1.5 Pro, Gemini Flash (multimodal, real-time web access)	High (Google Workspace for Education)	United States, North America
OpenAI	ChatGPT	GPT-3.5 (free), GPT-4, GPT-4o, GPT-5 (Plus), with DALL·E 3, Code Interpreter, Vision, Voice	Low	United States, North America
Anthropic	Claude	Claude 2, Claude 3 Opus, Sonnet, Haiku (security-focused, large context windows - does want to use data for training purposes)	Medium	United States, North America
Mistral AI	LeChat	Mistral 7B, Mistral 8x7B (open source, European privacy-focused)	Average	France, Europe
Proton	Lumo	Various open-source models. Proton presents itself as the most privacy-protective and European online environment	High (although there is no agreement with VU Amsterdam)	Switzerland, Europe
DeepSeek	DeepSeek Chat	DeepSeek-V2, DeepSeek-Coder (focused on mathematics, programming, Chinese-English bilingual)	Very low	China, Asia
xAI	Grok	Grok-1, Grok-1.5, Grok-2 (multimodal, real-time X/Twitter integration, 1M token context)	Very low	United States, North America

Table 6.1 Developers of LLMs, their service, models, and a broad indication of compliance with Privacy and Security requirements based on the knowledge of the AI Competence Network of VU Amsterdam.

6.5 Agentic AI and Artificial General Intelligence

AI application systems continue to develop: in 2025 and 2026 especially around so-called Agentic AI. In addition, there is a great deal of discussion about the development of so-called Artificial General Intelligence.

6.5.1 What is Agentic AI?

Agentic AI refers to combinations of prompts and search mechanisms that can think through problems, make decisions, and take actions with minimal human guidance. Unlike regular generative AI, which simply responds to prompts, Agentic AI behaves more like an autonomous assistant. These systems use LLMs to reason, have access to external tools (such as search engines or databases—or your own files), and have memory to learn from previous experiences. Think of it as the difference between a calculator that only computes what you enter and a study partner who can break your complex assignment into steps, look up information online, remember what worked last time, and adjust the approach if something does not work.

p. 6–69

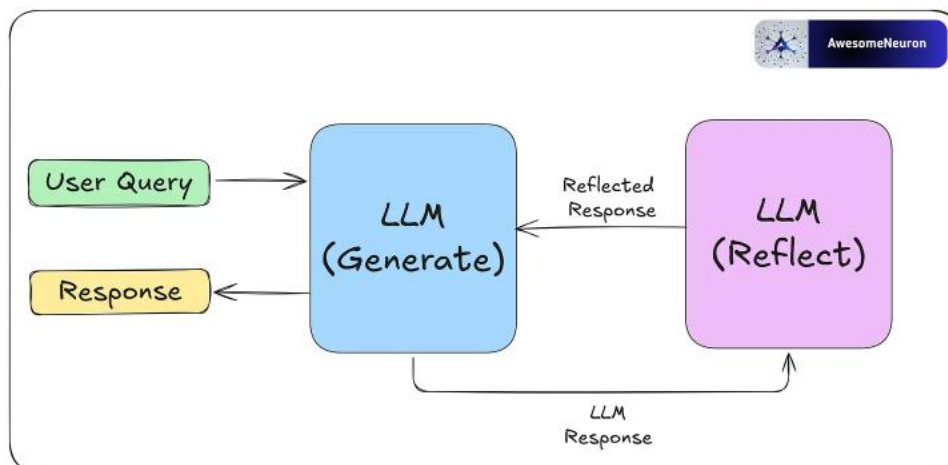


Figure 6.4 Illustration of a simple Agentic AI system. Prompts go back and forth between language models; each time the model responds to the input from the other model and sends it back again (until the decision is made to show the final result to the user. Image Tiwari, R.K. (2025).

If, for example, you asked an agentic AI to help solve errors in your programming assignment, it would not only propose a solution but also analyze the error, test possible solutions, learn from failures, write additional code for helper functions along the way, and improve iteratively until the code works (Herrman, 2026) (and it automatically keeps all the documentation up to date). But agentic AI can also be linked to your email program, electronic calendar, and files on your computer, so that it can also act on them directly such as automatically scheduling appointments or replying to emails. Another example is how AI-driven scientific literature search engines work. Although this makes them incredibly powerful for tackling complex, multi-step solution paths, they are not always necessary. For simple tasks, traditional AI chat methods are faster and cheaper.

6.5.2 Artificial General Intelligence (AGI)

Contemporary AI excels at specific tasks such as language processing, image recognition, or playing games. But Artificial General Intelligence (AGI) would mainly possess flexible, generalizable intelligence

capable of reasoning about unfamiliar problems, transferring knowledge across domains, and adapting to entirely new situations without additional training. Whether AGI will become reality is a much-discussed question in technology and science. Some researchers believe we may achieve it within decades by scaling up current approaches or through fundamental breakthroughs, while others argue that it requires discoveries we have not yet imagined, or that it may not be possible at all. The timeline and feasibility depend on unresolved questions about consciousness, reasoning, and whether intelligence can arise purely from computational processes, making AGI at once a technical challenge, a philosophical question, and a topic that requires careful consideration of its societal implications.

6.6 Data and privacy: who decides what happens to your data?

Providers of chat interfaces determine what happens to your data and the generated responses. The provider may, for example, choose to:

- store the chat history locally or in the cloud;
- use your data for analysis, product improvement, or even for training new models;
- share data with third parties or explicitly protect it.

p. 6-70

These policies differ from provider to provider and are often not fully transparent. Some interfaces offer privacy settings; others do not. That is why it is important to understand the technology described above and, based on that, make judgments such as:

- Are you working with a provider that has unclear terms and processes? Then be very careful with the data you upload and the questions you ask. Make sure you never share sensitive data such as personal information or confidential business information. Examples of this kind of provider include ChatGPT, Google, Anthropic, and many smaller providers. In other cases as well, the terms are often unclear, such as with a paid account or accounts that supposedly work based on signing in with your organization account.
- Does the university or your employer provide a chat environment, and are there clear agreements that the data will never be shared with parties outside the institution? Then you can use the chat environment much more freely, but still do not share sensitive data if it is not strictly necessary. Examples include the Copilot environment used by universities, at VU Amsterdam the Gemini functions, the SURF EduGenAI environment, or environments offered and managed by research institutes themselves. Note: login with your VU Amsterdam credentials.
- If the university or employer does not provide an environment, or if you doubt the agreements with the provider, then see whether you can work with an LLM that you can install on your own (powerful) PC or laptop and use through a local program (Local LLMs). Examples of programs include Jan.ai, LM Studio (Google), and Ollama.

Even so, a great deal often remains unclear about how your data is processed, and many legal and technical issues play a role. For more background on the complexity surrounding ownership of user-generated data in AI interactions, we refer you to the work of Metcalf & Crawford (2016).

Box 6.3 - Uploading files and data security

Are you allowed to upload files into chat systems? Assume that with providers of these systems that are free, you are actually giving those documents to those companies. They can reprocess that data (for model training) or even sell it, for example. With paid versions, there may be differences.

In addition, by uploading the files you may be infringing the copyright of the authors or publishers of the documents. You might think: “So what? Who is going to catch

me!” But that is exactly the situation in which you should rely on your own conscience and integrity. So when it comes to files containing special or sensitive data, you need to be especially careful.

You can prevent data leakage only if you use systems for which your institution has concluded a data processing agreement (at VU, for Copilot, Google Gemini or Gemini Notebook, or EduGenAI—with VU login) or if you use only Local LLMs.

6.7 Self-study questions

6.7.1 Check questions

1. What is the difference between an LLM file and the chat interface you use to communicate with the model?
2. How does 'conversational memory' work, and why does it seem as though an AI model remembers your earlier messages?
3. What factors should you consider when choosing among different AI providers such as OpenAI, Microsoft Copilot, or EduGenAI with regard to safety and privacy?

p. 6-71

6.7.2 Reflection questions

4. Suppose you upload a document containing sensitive information to ChatGPT for a study assignment. What risks do you run, and how could you approach this differently?
5. Many students think they are speaking directly with an AI model, but in reality everything goes through an intermediary layer run by a company. How does this affect your trust in and use of AI chat services?

6.7.3 Suggested answers

1. An LLM is essentially one large computer file that you could theoretically place on your own computer. It has no built-in chat functionality and is 'stateless' - it remembers nothing from earlier conversations. The chat interface is the intermediary layer (such as the ChatGPT website or app) that creates the conversation experience by storing the chat history and sending the entire conversation back to the LLM with each new question.
2. The LLM itself has no memory. The chat interface stores the full chat history and, with each new question, sends the relevant context back to the model again. This is called 'conversational memory'. This makes it seem as though the model remembers the conversation, but in reality it is seeing the whole conversation again each time.
3. You should look at factors such as transparency about what happens to your data, whether your institution has concluded data processing agreements, where the data is stored, whether your data is used for model training, and whether you can share sensitive information. EduGenAI via SURF and Copilot through your university license are usually safer than direct commercial services.
4. If you upload sensitive information to ChatGPT, it may be stored, analyzed, or used for model training by OpenAI. You run the risk that confidential data may end up with third parties. Better alternatives are to use your university's AI environment, install local LLMs, or remove the sensitive parts from your document before uploading it.
5. This knowledge should make you more aware of the choices you make. You realize that companies such as OpenAI, Google, or Microsoft determine what happens to your conversations. This may motivate you to choose more consciously for systems your institution

has made agreements about, or to be more critical about what information you share through commercial chat services.

7 How does your AI use impact the environment?

Generative AI uses a great deal of energy and water. And demand for the use of generative AI is increasing very rapidly, and so is demand for energy. Why is so much energy and cooling needed, and how does the climate impact of generative AI compare with the impact of, for example, industry, transportation, or agriculture?



p. 7-73

In this chapter you will learn

- Why training and using language models requires a great deal of energy and water.
- That the total energy use of data centers (of which AI and generative AI make up a large part) is rising exponentially and is expected to account for 10% or more of all energy consumption worldwide.
- How generating energy from coal and gas causes CO₂ emissions and how that warms the Earth.
- That Jevons paradox explains why demand for energy for AI is rising exponentially even though performing digital computations is itself becoming more efficient.
- How the use of generative AI compares with other services or products in terms of energy and water consumption.

Takeaway

- That you know that demand for energy and water for data centers is already high and is increasing very rapidly.

Keep in mind: that daily use of ‘ordinary’ services or products (e.g., watching Netflix, flying, or eating hamburgers) also uses a great deal of energy and water.

What do you think?

One hundred fifty years ago, very little energy was used worldwide, and in the decades that followed, the increase in energy use by industry, agriculture, and transportation apparently did not really lead to a problem yet (although warnings were already being issued in the 1970s that this could not continue). Why, then, do we see the use of energy and water specifically by AI as a problem? What do you think? What might be at play here?

Imagine ...

You ask ChatGPT a question or have Midjourney generate an image. But your fellow students tell you this is bad, because it uses a great deal of energy. If you follow the news and social media, it is full of this as well. You wonder: should I really use generative AI if it is so bad for the environment? And: how bad is it really?

7.1 What does your AI use really mean in terms of energy, CO₂ and water?

If you ask a generative AI a question through your laptop or phone, you do not see everything that is happening behind the scenes. Without realizing it, you are then using a large network of data centers, powerful graphics processors, and cooling systems that use a great deal of energy and water. In this chapter, we explore how this relates to environmental impact.

We look at three things: how much energy it takes to train models, how much energy and water are needed to generate answers, and how much CO₂ is emitted. We take a global perspective in doing so. Then we zoom in on your own use and compare it with the environmental impact of other everyday activities. Think of watching YouTube videos, flying to your vacation destination, eating a hamburger, or taking a shower.

p. 7-75

7.2 Why does training AI models require so much energy, water, and CO₂?

Training models requires a great deal of energy because the values of the billions of parameters in the neural network of the LLM have to be calculated. To do this, tens of thousands of processors (GPUs) from the company Nvidia are generally used. Each processor requires about 700 watts of power. You can compare this to a consumer microwave running at nearly full power. A great deal of water is needed for cooling, and energy generated from fossil fuels causes CO₂ emissions.

At the same time, it is difficult to determine how much energy is needed to train language models and what the CO₂ emissions are. First, because many factors play a role, such as the complexity of the model itself, but also the hardware it runs on, the questions you ask, and the energy source you use (for example, if you were to use renewable energy). Second, because the AI companies do not report their energy use (Vries-Gao, 2025b). In addition, the figures change quickly because of rapid technological developments. Much of the research on the environmental burden of generative AI is therefore based on estimates.

But how much energy are we talking about? Training runs for the current generation of LLMs, comparable to training sessions for GPT-4o, last about three months and consume about 45-56 gigawatt-hours of energy (You, 2025). So the training sessions require both high power and a long duration. The total amount of energy for training is therefore equal to the amount of energy needed to supply electricity to about 60,000-74,000 Dutch households for a year. That is quite a lot. It is sometimes claimed that you would even need a nuclear power plant for the energy consumption of a single data center. With the output of an average nuclear power plant (3000 MW), you could, for example, train 120 LLMs at the same time. If you use offshore wind turbines (10-15 MW each), you would need two or three of them. So comparisons should be made with caution.

So model training requires a great deal of energy, but how does that actually translate to a prompt? ChatGPT processes about 1 billion prompts per day (Ahmed, 2025). Suppose an LLM is used for about a

year and we spread the energy for model training across that number of prompts, then we arrive at about 0.151 Wh (watt-hours) per prompt. You can read more about that below.

7.3 How much energy does it take when you ask an AI a question?

Every time you generate text, an image, or a video with a generative AI system, servers are running in a data center. That uses energy - but exactly how much?



p. 7-76

Figure 7.1 Image of a typical data center: a very large enclosed building with many computers inside it (image source <https://www.northcdatacenters.com/>).

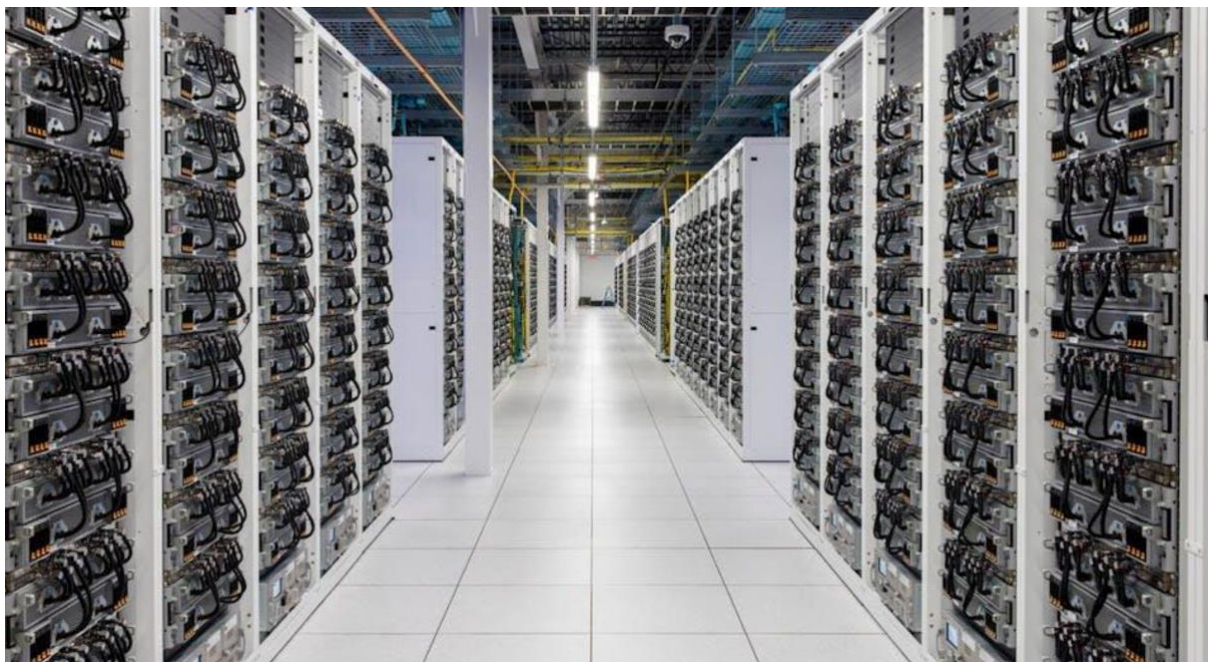


Figure 7.2 Interior of the xAI data center Colossus in Memphis, Tennessee. Colossus contains about 100,000 Nvidia Graphics Processing Units (GPUs) to train language models. Expansion to 200,000 GPUs is already being planned

(image source: <https://www.datacenterfrontier.com/machine-learning/article/55244139/the-colossus-ai-supercomputer-elon-musks-drive-toward-data-center-ai-technology-domination>).

Little is known about the precise energy consumption. Comparative research is done using open-source models (see section 9.7 for more explanation of this). But the most widely used models are closed and therefore cannot be tested. In addition, they change continuously: new, larger models and so-called ‘reasoning models’ (models that converse with themselves) often require more computing power and therefore probably use more energy.

To still get an impression, a comparison is often made between asking Google a question and giving an AI a prompt. This is a complicated comparison because little is known about Google's energy use as well. But for the sake of the discussion, we will look at it briefly. The energy use for a Google search is estimated to be around 0.3 watt-hours (Wh), for example. An estimate for an average ChatGPT request was around 2.9 Wh for earlier models - almost ten times as much (You, 2025). But thanks to technological advances and more efficient models, such as GPT-4o, the energy use per prompt has now fallen to about 0.3 Wh. That would therefore be comparable to a Google search (You, 2025). For a long input, however, the energy cost is higher: it can rise to about 2.5 Wh for an input the length of a magazine article, up to 40 Wh for an input of about 200 pages of text. Generating images and videos requires considerably more energy.

p. 7-77

7.4 How does prompting compare with other everyday energy use?

Whether the amount of energy needed to process a prompt is a lot or a little therefore depends on how you compare it with the use of other devices or services over a certain period of time. For example, 0.3 Wh is less than the amount of electricity an LED lamp or a laptop uses in a few minutes. Or equal to 0.9 seconds of use of a 1200-watt hair dryer or 7.2 seconds of operation of your 150-watt refrigerator compressor. So even for an intensive user, ChatGPT's energy costs will be only a small fraction of their total electricity use, and therefore also of the total amount of energy used worldwide. In the table below, we compare the amount of energy for 100 prompts (if you make intensive use of a generative AI system for a day) with the energy use of other everyday energy needs. Let the numbers sink in.

Device/Category	Power	Energy use	Wh per day	Time for 0.3 Wh	Time for 30 Wh
Hair dryer	1200 W	-	-	0.9 seconds	1.5 minutes
Refrigerator compressor	150 W	-	-	7.2 seconds	12 minutes
Laptop	50 W	-	-	21.6 seconds	36 minutes
LED lamp	10 W	-	-	1.8 minutes	3 hours
100 ChatGPT prompts	-	30 Wh total*	30 Wh*	-	Reference point
100 ChatGPT prompts per day	-	10.95 kWh/year	30 Wh	1 day**	1 day

2 hours of watching YouTube per day (tablet)	22 W	16.06 kWh/year	44 Wh	49.1 seconds	1.36 hours
2 hours of watching Netflix per day (TV)	115 W	83.95 kWh/year	230 Wh	9.4 seconds	15.7 minutes
Average Dutch household	331 W	2900 kWh/year (CBS, 2018)	7945 Wh	3.3 seconds	5.4 minutes
Standby consumption of a Dutch household	51 W	450 kWh/year (United Consumers, 2025)	1224 Wh	21.2 seconds	35.3 minutes

Table 7.1 A comparison between the energy required for 100 generative AI prompts per day and other everyday energy use.

7.5 What does your AI use mean in the broader energy picture?

p. 7-78

Although the energy use per prompt is small, the impact on a global scale becomes clear when you multiply it by the millions of daily prompts. In 2023 alone, ChatGPT's annual energy use was estimated at 226.8 GWh - enough to fully charge more than 3 million electric cars once (You, 2025).

According to a report by the International Energy Agency (Cozzi & Gould, 2024) the processing of all digital data, mainly for AI, will consume more electricity in the US alone by 2030 than the production of steel, cement, chemicals, and all other energy-intensive goods combined. According to the report, global electricity demand from data centers will therefore more than double by 2030. It could possibly rise to an energy use amounting to 10 percent of total energy consumption in the United States. AI will be the main driver behind that increase, with demand from dedicated AI data centers alone expected to more than quadruple. One data center currently uses as much electricity as 100,000 households, but some data centers currently under construction will need twenty times as much electricity.

Newspaper headlines recently suggested that AI will cause a 165 percent increase in data center electricity use by 2030 (Verhagen, 2025). Of the increase to 650 TWh, general internet growth and general AI applications will contribute about 200 TWh. However, that increase had probably already been reached by the end of 2025 (Vries-Gao, 2025b).

It is important to keep in mind here that general AI use is much broader than generative AI applications alone, according to Masley (2025a), an essayist who takes a critical look at AI doom scenarios (see the note later in this chapter). Examples of general internet and AI use include:

- powering recommendation systems for the personalization of, for example, streaming platforms, e-commerce sites, social media feeds, and online ads.
- business operations such as data analysis, predictive and decision-making processes, search, and ad personalization.
- computer vision such as object detection, facial recognition, video and medical image analysis, and content moderation for automatically flagging inappropriate images.
- sound and audio (think of voice assistants and speech recognition systems, such as Amazon's Alexa, Google Assistant, and Apple's Siri).

Chatbots such as ChatGPT account for only 3 percent of total AI use (or 0.6 percent of total data center energy use). So that is a limited share of the energy landscape, but it is growing; see also Figure 7.3 (Masley, 2025a).

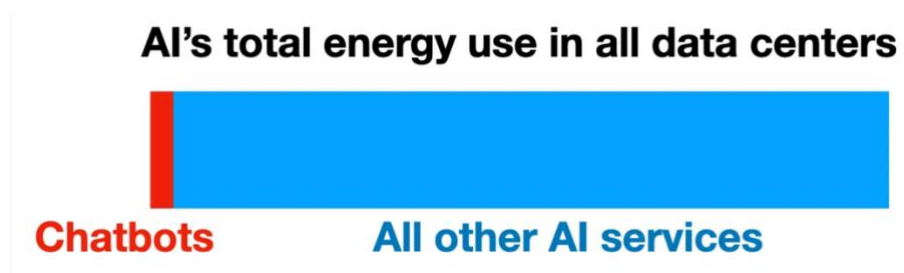


Figure 7.3 Comparison of the amount of energy chatbots consume compared with all other AI services (bron grafiek: Masley, 2025a).

For comparison, YouTube currently uses about 1 percent of global energy - about ten times as much as the expected use of chatbots in 2030. Concerns about ChatGPT's energy impact are therefore probably best weighed proportionally against other digital services.

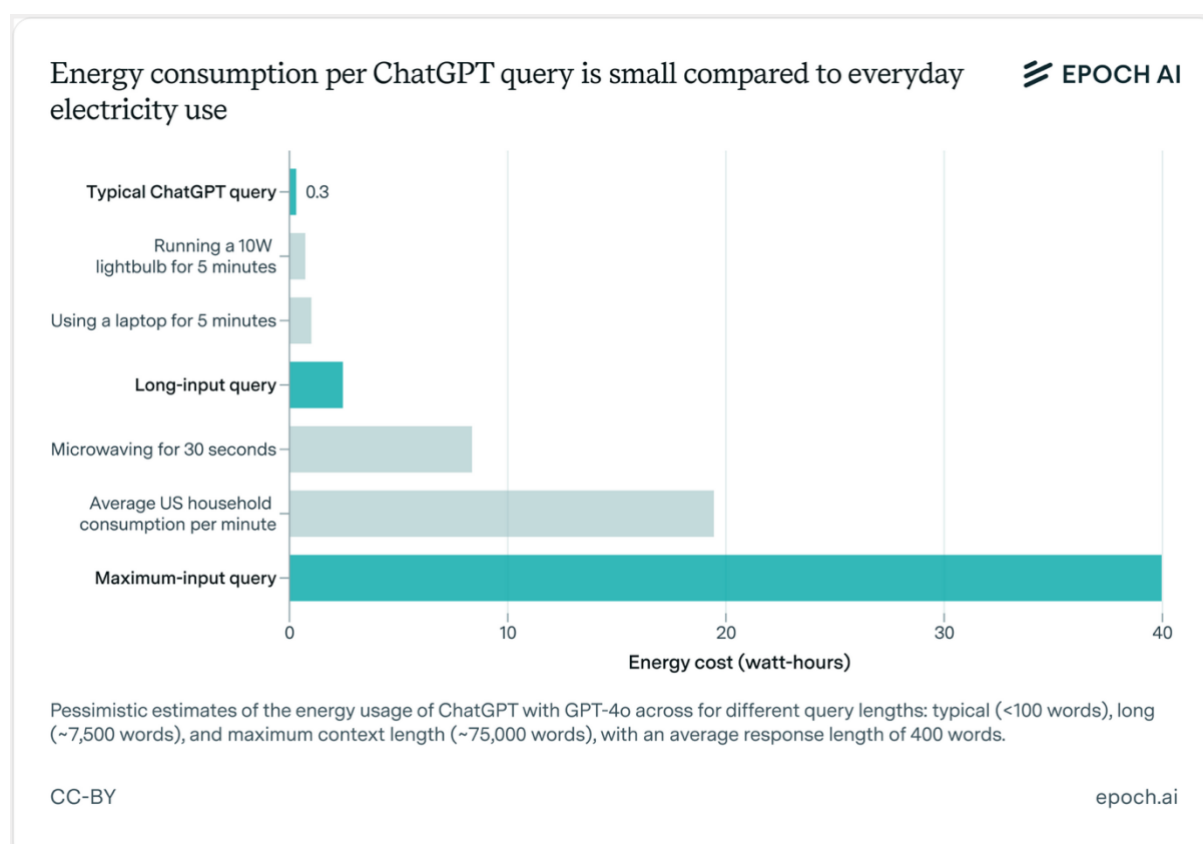
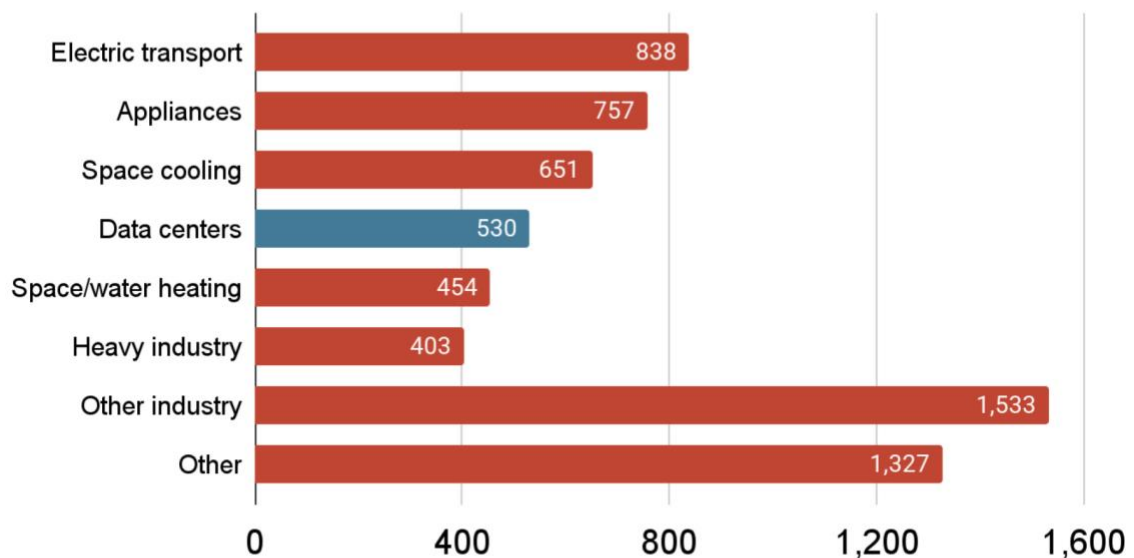


Figure 7.4 Comparison of energy use with other forms of everyday energy consumption (bron grafiek: You, 2025).

Want to know more? Kai Williams (Williams, 2023) has written an article that uses 16 graphs to show what the growth of AI means.

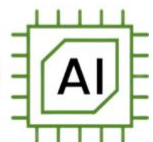
Other power demand growth could dwarf AI

Projected growth in electricity demand in TWh, 2024-2030



Source: International Energy Agency,
Energy and AI, 2025

UNDERSTANDING



p. 7-80

7.6 CO₂: How AI contributes to global warming

In the future, the energy needed for data centers could come from green and renewable sources - and a great deal of work is being done on that (OECD, 2022). But at the moment, this energy mainly comes from burning fossil fuels: coal and natural gas. This releases CO₂ emissions. And although AI causes only part of the total amount of CO₂ emissions, it is still important to understand the effects of increased CO₂ concentrations in the atmosphere in order to properly assess the sometimes heated debates about it in the media, on social media, and in politics. This involves two important effects of increased CO₂ concentrations in the air and their consequences. These are described in a compelling way by Andri Snær Magnason in the book 'On Time and Water' (2022). Below you can read more about those insights.

7.6.1 Global warming and its consequences

First, increased CO₂ concentration causes global warming. Without additional climate policy, there is a risk that by around 2050 we will be heading toward 2 °C of warming. That may not seem like much, but it is. Especially because the effects differ by region and continent (and can therefore also be much greater):

- Large areas are already becoming uninhabitable (and even more so in the near future) because it is simply too hot to live there.
- More severe periods of drought and heavy rainfall disrupt crop growth.
- There will be more and larger forest and land fires.
- There will be more and larger floods.
- There will be more erosion and more fertile soil will wash away.
- The weather will become more extreme and fluctuate more unpredictably.
- More flooding and greater risks for areas that could end up under water.

All these phenomena have major consequences. Let us zoom in briefly on the accelerated disappearance of glaciers as an example. One medium-term effect is, for example, that areas whose residents depend on the water supplied by the rivers traditionally fed by these glaciers become uninhabitable. Think of densely populated areas in Asia (from rivers fed by glaciers in the Himalayas) and Latin America (rivers fed by glaciers in the Andes). This will cause large-scale population migrations. The melting of glaciers also raises sea levels. As a result, large low-lying coastal areas will become saline or be submerged. On the one hand, this creates problems for growing crops; on the other hand, it will set migration flows in motion. In addition, higher temperatures will cause the vast global tundras (large areas bordering a polar region and consisting of grasses, mosses, lichens, and dwarf shrubs) to thaw slowly. This releases large quantities of methane into the atmosphere. Methane has an even greater effect on retaining heat on Earth than CO₂. This further accelerates global warming.

7.6.2 Ocean acidification and its consequences

Second, increased CO₂ concentration causes higher acidity in the oceans. CO₂ is absorbed by the oceans, but by now that amount is so great that it substantially changes the pH level of the water. As a result, calcium carbonate formation for organisms (shellfish and reefs) slows down or even stops, causing all life that feeds on these organisms to die out. This will have an enormous effect on biodiversity in the oceans and will also affect our food production through fisheries.

p. 7-81

A great deal of work is being done to reduce CO₂-emissions (for example, by capturing them directly at the source) or to capture CO₂ from the atmosphere or seawater. However, these techniques are still in the early stages, and the scale needed to substantially lower the percentage of CO₂ in the atmosphere is enormous. It would require an international approach and unprecedented investment.

7.6.3 Do you substantially reduce CO₂-emissions by not using generative AI?

The demand for electricity will increase in the coming years. This does not necessarily have to lead to higher CO₂-emissions if governments and industry commit to a clean energy mix: from wind farms and solar fields to even nuclear power plants (although these, in turn, pose a waste problem in the long term).

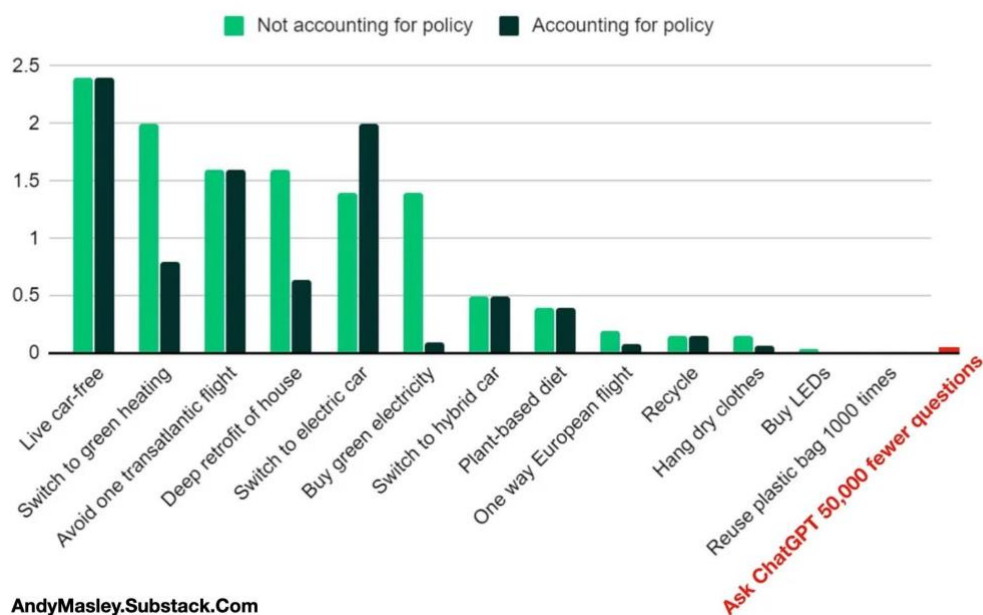
However, as long as the world continues to rely heavily on burning fossil fuels to meet its energy needs, this will indeed lead to global problems. Many journalistic, social media, and even scientific sources argue that we shouldn't use generative AI at all because of the environmental impact. And likely due to the novelty, speed, and scale of growth in the use of AI—and generative AI in particular—this point is heavily emphasized in climate discussions. How effective is that?

By placing other forms of CO₂ emissions alongside AI use, you get a more realistic and broader perspective on the relationship between them. For that, Masley (2025a) analyzed changes in lifestyle and consumption patterns for their effects (See Figure 7.5). Masley shows how much CO₂-emissions you can avoid by making other lifestyle decisions, compared with 50,000 fewer ChatGPT prompts. This shows that the share of ChatGPT prompts pales in comparison with other lifestyle changes such as no longer driving, skipping a flight once, becoming vegetarian, or hanging your clothes to dry instead of using the dryer.

Masley therefore wonders whether we would do better to focus our efforts on those factors instead of placing one-sided emphasis on the environmental burden of generative AI.

It is good to know who Masley is when you read Masley's figures. Masley is not a scientist who publishes peer-reviewed research, but an essayist who shares analyses on a personal blog. Masley takes a clear position in them and wants to correct the idea that generative AI would be a major environmental problem. One of these pieces even carries the firm title that AI's water problem would be "fake." That does not make the work unusable, because the comparisons are traceable and the underlying data are often verifiable. But it does mean you should read the conclusions as an argued position rather than as a neutral scientific finding. A good exercise in AI literacy is therefore to ask yourself: what choices does Masley make in these comparisons, and would a different choice lead to a different picture?

Figure 4. Tonnes of CO₂ avoided by selected personal lifestyle decisions accounting for government policy



p. 7-82

Figure 4. Tonnes of CO₂ avoided by selected personal lifestyle decisions accounting for government policy

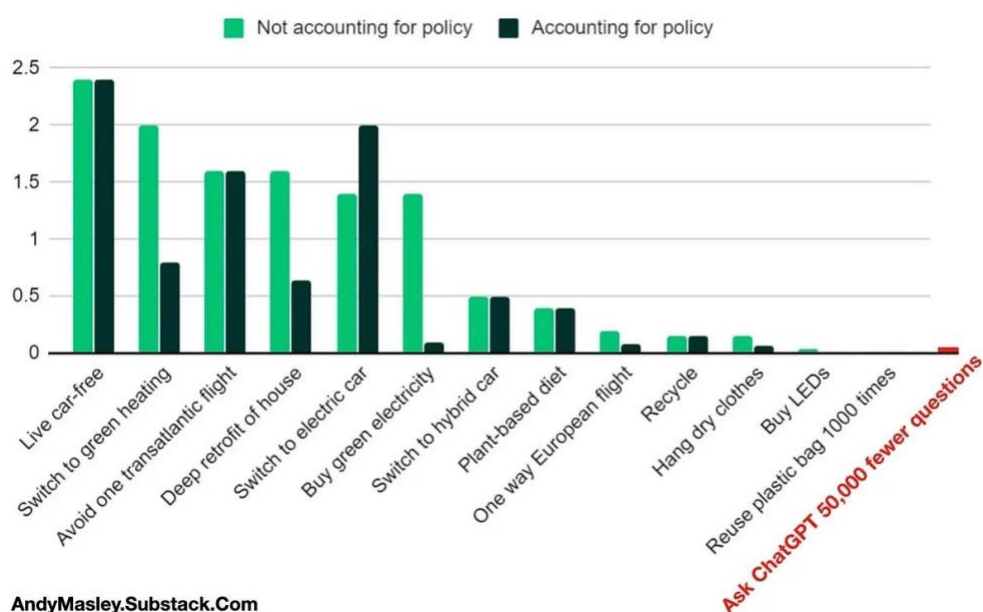


Figure 7.5 Graph showing which lifestyle change results in a certain reduction in CO₂ emissions (in tons) (bron grafiek: You, 2025).

7.7 Water: why your AI gets thirsty too

7.7.1 Global and local water use

You may not realize it when you type in a prompt, but AI systems also use relatively large amounts of water (You, 2025). Servers in data centers generate a great deal of heat and therefore need to be cooled, currently still with water. On a global level, the impact of this is significant. Tech companies such as Microsoft, Meta, and Google together used about 2.2 billion cubic meters of water—comparable to the annual water consumption of all of Denmark. And this consumption is growing rapidly, with global demand in 2027 expected to amount to 4.2 to 6.6 billion cubic meters of water withdrawal (Li et al., 2025).

The biggest problem is not even the total scale of the water use (the water returns to the cycle), more on that later, but rather that it causes local problems: data centers are also being built in areas where water scarcity already exists. This leads to competing interests between the need for cooling water at data centers and water use for agriculture, industry, and households.

Mesa, for example, is a city in a desert region of Arizona that illustrates this problem well (Solon, 2021). One important reason local governments want to host data centers is jobs. And a desert region with lots of solar and wind power offers data centers cheap electricity. Arizona therefore seems suitable for a data center. But its water needs were not taken into account. The data center built in Mesa in 2021 uses up to 15 million liters of water per day, even though the region has been dealing with extreme drought for decades. Altogether, data centers in Mesa used more than 3.8 billion liters of water per year in 2021: enough to supply tens of thousands of households with water. This leads to major local tensions.

p. 7-83

7.7.2 Water use per prompt

But how much water does generative AI actually use compared with other activities? In the media or online, you may often hear that a single prompt costs half a liter of water. But that is not correct: according to recent estimates, the amount of cooling water needed to process 10 to 50 prompts through ChatGPT is about half a liter of cooling water (Li et al., 2025). By comparison, Masley says that watching streaming videos in particular, for example, uses a great deal of energy and water. Masley also compares this with the water used to produce a hamburger (agriculture) or the total volume of water lost through leaks in the US. See Figure 7.6 and Figure 7.7. These considerations lead Masley to observe that the 500 milliliters of water used for twenty ChatGPT prompts can also be offset simply by showering five seconds

less or skipping a hamburger.

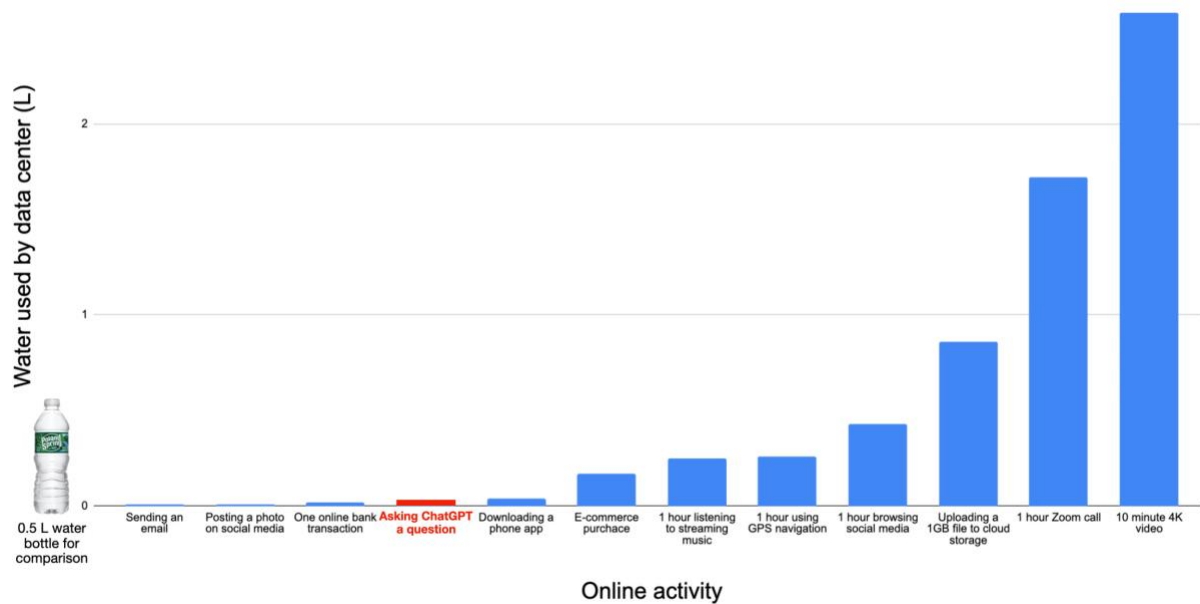
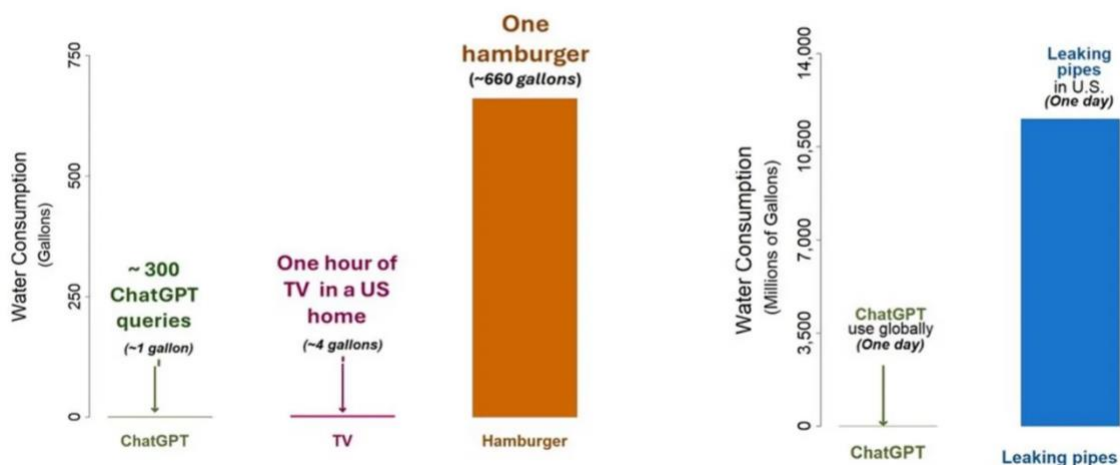


Figure 7.6 Comparison of water use by various online activities (bron grafiek: Masley, 2025a).

Water consumed using ChatGPT vs other activities



Sources: Li, Ren et al., 2023; U.S. Census Bureau; UNEP

Liemberger & Wyatt, 2019; Liemberger & Wyatt, 2020

Figure 7.7 Comparison 2 of water use by various functions (bron grafiek: Masley, 2025a).

The question is not so much whether data center water use is inherently bad, but rather whether deliberate choices are being made. In water-rich areas, water use for data centers is probably not a problem, but in areas with water scarcity, every drop is an ethical consideration. The question should be: is this the best location for a data center? Are there alternatives for cooling? It is currently being investigated whether building data centers in space (“SpaceX neemt xAI over en wil datacenters in de ruimte bouwen,” 2026) or at sea, directly next to wind farms (“Vattenfall en Project Enki willen datacenters bouwen bij windparken op de Noordzee,” 2026), may be possible solutions. And how do the economic benefits compare with the ecological and social costs? Doesn't space become polluted? The water use of data centers is relatively small compared with agriculture, but in the context of extreme drought every liter counts.

7.8 When does AI actually help the environment?

So is AI always bad for the environment? Not necessarily. In a variety of sources, we find that AI applications can actually contribute to sustainability (Stern et al., 2025). Think of:

- optimizing energy use in buildings;
- faster analysis of climate data;
- predicting agricultural yields;
- planning logistics and mobility more efficiently.

An example: Researchers from DeepMind, Google, and Trane Technologies had an AI system control the cooling systems of two large buildings for three months: a university campus and a complex containing stores, residences, and medical facilities. The system learned from sensor data and alternated daily with the standard control software. This resulted in energy savings of approximately 9 and 13 percent (Luo et al., 2022). The system continued to improve throughout the trial, and no new hardware was required.

Another example is Project Green Light. In this initiative, Google analyzes anonymous traffic data from Google Maps. Based on this data, cities receive recommendations on how to adjust their existing traffic lights. In the first twelve participating cities - including Hamburg, Manchester, and Seattle - the modified intersections saw up to 30 percent fewer stops and up to 10 percent lower emissions (Matias, 2023). No new infrastructure is required, making this approach viable even for cities with smaller budgets.

There is also serious opposition. Alpine et al. (2026) use a global energy-economy model and find that, on balance, productivity gains from AI lead to 0.47 to 1.8 Gt of additional CO₂ per year. The reason is that AI makes the fossil fuel sector at least as efficient as the renewable sector and thus encourages the use of fossil fuels. Furthermore, most of the positive figures are potential estimates or results from small-scale trials. Stern et al. and the IEA work with scenarios. The cooling trial was conducted at two locations. The figures from Green Light are “up to” values from a Google corporate blog and have not been independently verified.

Using AI for these kinds of applications is very complex, which makes it difficult to determine to what extent positive results actually come from them.

7.9 The information-energy paradox

Another perspective concerns the relationship between the energy use of digitization and generative AI and the amount of information it yields (Masley, 2025b)? If you put the energy use of digital information sources into perspective, a striking picture emerges. For 0.3 watt-hours—the amount of energy a ChatGPT prompt uses—you get, according to Masley, access to a synthesis of billions of documents' worth of knowledge. By comparison, looking up the same information in a physical library would take hours, including travel to the library, the lighting of the building, and possibly the printing or copying of sources. The shift from physical to digital information carriers has produced an enormous environmental gain that we often overlook, although it had already begun before the arrival of generative AI. A scientific journal that used to be mailed monthly to thousands of subscribers worldwide—with all the paper production, ink, transportation, and waste processing that entailed—is now accessible through a download of a few megabytes. Even if we include all server costs and individual energy use, the CO₂ footprint per article consulted has dropped drastically.

In addition, according to Masley, generative AI has the ability to multiply knowledge without a proportional increase in energy use. Once an AI model has been trained, it can serve millions of users at the same time, with each additional user adding only marginal cost. An AI-generated summary of

complex scientific literature that would previously have required days of research is now produced in seconds for the energy cost of seven seconds of blow-drying your hair. This gives every student worldwide access to the same advanced AI tools without the environmental costs of international conferences, study visits, or shipping study materials. In that sense, the knowledge yield per watt-hour of modern AI systems is unprecedented—an efficiency that earlier information systems never came close to.

Still, something is unsettling. And that has to do with the scale at which everything is happening and the fact that new techniques—if they are cheap enough—lead to a multitude of applications that were not initially foreseen. The Jevons paradox explains this.

7.10 Where is this going? The Jevons Paradox

Even because of (or precisely because of) the fact that models and techniques are becoming more efficient, the total amount of energy required to train and use generative AI systems is growing.

How is this possible? An explanation is provided by the so-called Jevons Paradox. That theory does not predict much good (Polimeni et al., 2015): efficiency gains in (sub)systems are always accompanied by large growth in the energy demand of the overall system because that efficiency also makes more use cases and outputs of systems possible. And that certainly applies to AI and generative AI as well (Luccioni et al., 2025). Because the possible applications of generative AI - because it is relatively inexpensive - are growing at a dizzying pace and can be made available to vast numbers of users at almost no additional cost. In other words, cheaper and more energy-efficient digital processors actually drive increasing demand for digital generative AI services, and with it the energy demand.

At the same time, it is important to also consider the relative contribution of generative AI to environmental impact compared with other activities, and the savings from the paradoxical perspective in paragraph 7.9 in your considerations. Ask yourself: is there a provider that uses renewable energy sources? For example, data centers that you know run on solar or wind power. And don't just look at the costs of using AI—also factor your goal into your decision-making. You want to look something up or create something, but is generative AI really necessary for that? Don't just ask, "How much energy does this consume?" but also, "What value does this provide?" Critical use of AI is then no longer the result of a cost analysis, but of a—more comprehensive—cost-benefit analysis. What do you think?

These two sides—the nuance offered by Masley and the warning conveyed by the Jevons paradox—do not have to exclude each other. Data researcher Hannah Ritchie describes in the book *Not the End of the World* (2024) how you can put facts into perspective without slipping into indifference. Ritchie's stance can be summarized as "urgent but not self-satisfied": the figures show that many doom scenarios are exaggerated, but that is a reason to act purposefully, not to sit back. Note that Ritchie writes about the environment in general, not about AI, and describes an attitude rather than something about AI. For your AI use, this means that you can put the environmental impact of a single prompt in perspective, while at the same time continuing to take seriously the fact that AI's total energy and water use worldwide is growing substantially and calls for real choices.

As with other technologies, the key lies in responsible use and careful consideration:

- Don't use AI just because you can; use it because it actually improves the desired outcome.
- Choose consciously whether to use devices or services less in general—especially airplanes and cars.
- Choose a different lifestyle to reduce strain on energy resources and the environment; for example, become a vegetarian or flexitarian, watch fewer YouTube videos, and limit your use of social media.

- Choose consciously whether or not to use AI applications in relation to the information-energy paradox and the Jevons Paradox.
- Even if you use generative AI and find it useful, support any (future) regulatory requirements aimed at reducing its use.
- Don't just ask, "How much energy does this consume?" but also, "What value does this provide?"

7.11 Self-study questions

7.11.1 Review questions

1. According to the book, how much energy does an average ChatGPT prompt use compared with a Google search?
2. Why do AI data centers need water, and about how much water does it take to process 20 to 50 simple prompts?
3. Name two ways AI applications can actually contribute to sustainability.

7.11.2 Reflection questions

4. Suppose you use about 50 AI prompts every day for your studies. How would you weigh this energy use against other choices you could make to reduce your ecological footprint? Explain how you would make that trade-off.
5. Ritchie (2024) argues that looking at data soberly actually creates room to act. Do you find that convincing, or do you run the risk that putting things in perspective turns into looking away? Support your answer. Does this make you think differently about when you do or do not use AI?
6. Listen to [this podcast](#) from Trouw and try to formulate some criticism of the podcast's apparent message.



Figure 7.8 Podcast from Trouw, June 25, 2025.

7.11.3 Suggested answers

1. An average ChatGPT prompt uses about 0.3 watt-hours, which is comparable to a Google search. Earlier models used about 2.9 watt-hours, almost ten times as much, but this has dropped sharply due to technological progress. At the same time, models are becoming more powerful, which requires more energy. But processors are becoming more efficient again. That in turn requires more energy. Basically, nobody knows exactly where this is headed.
2. AI data centers need water to cool the servers that generate a lot of heat. About 500 ml of water is needed on average for roughly 20 to 50 simple prompts, depending on the model, the prompt length, and the location of the data center.
3. AI can contribute to sustainability through, for example, optimizing energy use in buildings, faster analysis of climate data, predicting agricultural yields, and planning logistics and mobility more efficiently.
4. 50 prompts per day cost about 15 watt-hours, which comes to 5.5 kilowatt-hours per year. This is comparable to showering about 5 seconds less per day or watching YouTube for a few fewer

minutes. You could weigh this against the fact that other choices, such as flying less, eating vegetarian, or driving less, have a much greater impact on your ecological footprint.

5. This is a personal reflection in which you might consider that the information makes you more aware of the relative impact of different activities. You might conclude that you can still use AI responsibly for meaningful purposes, but perhaps become more aware of excessive use. Or you may realize that other lifestyle choices matter more for the environment than limiting AI use. Or that you want to raise the issue of energy waste within your organization or through laws and regulations.
6. The podcast seems to endorse the general tendency to say that generative AI is bad for the environment. But in making that point, everything related to digitization is lumped together. You can argue that this improperly exaggerates the discussion about the impact of generative AI.

7.12 Activity - AI Environmental Challenge

Test yourself on your environmental impact.

<https://canvas.vu.nl/courses/83333/pages/activiteit-ai-environmental-challenge>

p. 7-88

7.13 Activity - Where do you stand on the environmental impact of generative AI?

If you have read chapter 7, you will notice that the chapter makes it clear that there is a major impact on the climate and the environment. At the same time, there is concern that Europe could fall behind technologically and economically compared with, for example, the United States and China if we were to ban the technology (we will go into this in more depth in the next chapters). What is your opinion on this now that you have read the chapter? Share your opinion with your fellow students on [Canvas](#). But stay friendly and remain open to arguments.

8 Generative AI and work

Generative AI affects your future work. Where is more work likely, and where less? For which disciplines is this more true than for others? What kinds of skills will be important to do that work in the future?



p. 8-89

In this chapter you will learn

- That the nature of much work will change because of AI.
- That the changes can vary greatly by discipline.
- That access to AI and differences in knowledge can lead to major social or economic disparities.

Key takeaways

- AI usually does not replace entire professions one-to-one, but mainly changes tasks, roles, and expectations.
- Those who work with AI must not only learn to use it, but also learn to direct, check, explain, and justify it.
- The benefits of AI are distributed unevenly: access, skills, data, language, and power determine who benefits.
- Responsible AI use requires up-to-date knowledge about developments and an active attitude.

Keep in mind: generative AI is not a neutral tool; it changes work processes, power relations, and expectations.

What do you think?

Do you think it is your own responsibility or that of your program (or your future employer) to keep your knowledge about AI up to date.

Imagine

You are in your final year of study and starting to think about your future. You wonder what the job market will look like by the time you have your diploma in hand. What if generative AI does part of the work you are now being trained for? How do you prepare for a future in which AI plays a central role in almost every profession?

8.1 The changing nature of work

You may picture yourself as a physician, lawyer, teacher, or policy advisor. But what do the latest AI developments mean for the way you will collaborate, make decisions, or be held accountable? AI's role is not limited to a single field or profession. As a student, you are developing skills for the job market, but it is also important to learn to reflect critically on the role you yourself can play in the design, deployment, and oversight of AI systems within your future profession. Whether you become an engineer, policymaker, or care professional: your decisions will help determine how AI is used - or misused.

AI changes the nature of work not only through automation, but also by creating new roles and ways of working. So, while some jobs will disappear, new ones will also emerge—ones we can't always predict. This has also been the case throughout history with major innovations such as the introduction of the printing press, the computer, and the internet. According to McKinsey (2023) by 2030, a significant share of routine cognitive and physical tasks will have been automated. But at the same time, new professions are emerging that require collaboration with AI, such as an AI prompt specialist, an ethical designer, or a data curator.

Generative AI plays an important role in this. While traditional automation mainly replaces repetitive tasks, generative AI systems enable people to create content independently, make creative suggestions, and even provide policy support. Think of AI that helps HR staff write job postings, lawyers who use AI to analyze case law, or healthcare professionals who use AI to summarize patient records.

Changes in work processes are at least as important as job titles. AI changes how we communicate, plan, decide, and collaborate. Work is therefore becoming hybrid: part human, part machine. This calls for new skills such as critical thinking, AI literacy, and digital ethics (McAfee & Brynjolfsson, 2017)

8.2 AI in different sectors: from healthcare to journalism

The impact of AI differs by sector. In journalism, AI can summarize news or write personalized articles. In the legal system, AI analyzes thousands of case files to discover patterns. In medicine, AI helps with diagnoses based on imaging. In the construction sector, AI-driven drones and robots can carry out inspections or generate models.

In higher education, you already see applications in which AI functions as a tutor, evaluates draft reports, or helps students formulate research questions. In the creative sector, collaborations are emerging between humans and AI for designing campaigns, music, or even architecture.

Listening tip: do you want to hear more about the opportunities and limitations of AI in medical practice? Listen to the podcast De Technoloog from BNR, episode: [AI in healthcare: breakthrough or overrated](#) (episode 430, 2024). Professor Henk Marquering discusses how AI is already being used in diagnostics, what it actually delivers, and why regulation is crucial.

8.3 Socioeconomic impact: inequality

AI changes not only how we work, but also what work means. If some tasks disappear, questions arise about income security, social status, and ownership of labor. And a digital divide emerges.

Box 8.1 - The digital divide in society

AI does not distribute opportunities equally across society. Groups that have little access to technology, understand how it works less well, or are less adept at using it risk falling behind. This is called the digital divide, and it involves more than simply owning equipment. Research distinguishes three layers: differences in access, differences in skills and use, and differences in what people ultimately achieve with it (Van Dijk, 2020).

The consequences affect two areas. In the labor market, people without AI skills have fewer opportunities for work and advancement, which reinforces existing socioeconomic differences. In addition, those who do not understand how AI and algorithms work are more vulnerable to directed information through digital and social media, such as misleading content or targeted political messages. The digital divide is therefore not only an economic issue, but also a democratic one.

p. 8-92

Who benefits from AI? Large companies with a lot of data, or also small and medium-sized businesses and freelancers? And what does that mean for the distribution of wealth? AI can disrupt economic models: think of platform economies that restructure entire professions, or 'data dividend' models in which citizens are compensated for the use of their data (Sadowski, 2019). These shifts have direct consequences for social inequality, labor rights, and democratic legitimacy. Moreover, ethical tensions are arising around access to AI-supported jobs. What if only people with the right digital or language skills can succeed in an AI era? What does that mean for diversity and inclusion in the workplace?

Box 8.2 - Ethics in the future of AI

AI raises fundamental questions about justice, autonomy, and public control. Who decides how AI is deployed in public services such as education, healthcare, or safety? How do we prevent AI decision-making from becoming opaque or undermining human dignity? Transparency, explainability, and participation are crucial. At the same time, we must remain alert to new forms of exclusion: for example, if AI systematically disadvantages certain groups in job applications or credit assessments. In democratic processes, there is a risk of AI-generated influence, microtargeting, or manipulation. The future of AI is therefore not only technical, but also moral and political (Binns, 2018; Eubanks, 2018).

That is why there are international initiatives advocating for an 'AI Bill of Rights,' which sets out fundamental rights and obligations around AI. You, as a future professional and citizen, also have a role in this as a co-architect of the AI future. The questions of who develops, deploys, and controls AI touch on broader power structures in society. In the next chapter, we therefore explore how AI is becoming intertwined with politics, media, business, and public values — and what that means for you as a critical citizen.

8.4 How do you keep up? Strategies for staying informed about AI

AI is one of the fastest-developing fields of our time. Knowledge that is current now may be outdated in two years. That is why it is important to train yourself in ongoing curiosity. Follow the news in professional journals and news programs and join networks where AI discussions are taking place.

Online platforms such as arXiv, Medium, LinkedIn, or Substack also offer accessible analyses of the latest trends, often written by researchers themselves. Do not forget the AI forums at your own university or student association either: more and more programs are integrating AI events and interdisciplinary projects. Keeping up with developments in AI is not a luxury, but a necessity. AI affects legislation, policymaking, technological standards, and public values. Those who stay informed also stay resilient. The skills to understand, evaluate, and use AI responsibly are becoming indispensable in virtually every profession.

8.5 Self-study questions

p. 8-93

8.5.1 Questions

1. Name two ways AI will change work in the future.
2. Why is lifelong learning important in an AI-driven labor market?

8.5.2 Reflection questions

3. Do you think it is your own responsibility or that of your program (or your future employer) to keep your knowledge about AI up to date.
4. Which professional group do you think will benefit the most from, or be harmed the most by, the arrival of generative AI?

8.5.3 Suggested answers

1. Through the automation of tasks and the emergence of new professions.
2. Because job roles change and new skills are needed.
3. Professional groups where manual work is important will be less affected than people who work, for example, in administrative or customer service departments.

9 Societal forces around AI

AI affects not only your work, but society as a whole. This ranges from those whose work is (illegally) used to train models to the impact of generative AI output and social media algorithms on politics or on the well-being of individuals. The owners of the technology can increasingly exert influence to impose their will on others.



p. 9-94

In this chapter you will learn

- What kinds of (illegal) sources language models are built on.
- What exhausting and dubious labor is performed to train language models.
- How algorithms hold your attention and amplify extreme messages.
- Some basic concepts related to privacy.
- What the growing influence of a few powerful AI and social media companies is.

Key takeaways

- Both the development and use of generative AI have societal effects.
- The influence of large AI companies on society and politics seems only to be growing. And that is not inherent to generative AI, but above all a deliberate choice by these companies.
- It is possible to escape the grip of the big tech companies by making deliberate choices for open source models and, for example, running language models locally.

Keep in mind: AI is not neutral, not only because of the way language models are trained and the sources they are trained on, but also because of how they are then deployed by large companies.

What do you think?

Should there be more rules to curb the influence of big tech companies on society? Or do you think it's not that bad?

Imagine ...

You read the latest news about AI developments and wonder: where does the data used by AI chatbots come from? Who actually sets the rules? What happens to my input? I often hear that AI is further polarizing society. But how does that really work? Can someone enter into a relationship with an AI chatbot?

9.1 Introduction

In this chapter, we discuss a number of societal aspects of generative AI. These include aspects related to the creation of language models, the way AI can polarize or instead create connections, and the power of companies that develop and offer AI techniques.

p. 9-95

9.2 Data sources and copyright issues surrounding LLMs

To build language models, large amounts of data are needed, but it is often unclear how the companies that make LLMs obtain that data. Previously, companies were fairly transparent about the data sources they used to train their models. A large portion of the Dutch texts used to train AI models, such as ChatGPT, for example, comes from the so-called Common Crawl database (mC4 dataset), which functions as a representation of the internet. De Groene Amsterdammer and Data School (Hofman & Veerbeek, 2023) analyzed these sources.

This analysis revealed three striking categories:

- The largest source of information in the dataset is the DocPlayer website. This website was set up by a Russian entrepreneur and automatically collects documents that appear online. The documents are often reused without permission, which makes this source legally and ethically problematic. In addition, the site also contains a great deal of personal and sensitive data. For example, tax returns and job application letters have been found there. That makes the use of this source particularly controversial.
- Other important sources are articles from newspapers and journalistic magazines. These are journalistically high-quality sources, but the publishers never gave permission for this use. The developers of LLMs make free use of data in which publishers have made substantial investments of time, money, and journalistic labor. Publishers are therefore exploring legal avenues to change this situation.
- Many of the sources on which LLMs are trained are in fact low quality, with their reliability seriously in question. For example, the reviews people leave in webshops. In addition, the Nazi conspiracy website Stormfront also turns out to be one of the sources used, and in terms of size in the C4 source list it ranks only one place below RTL Nieuws.

Welke Nederlandse websites worden gebruikt door chatbots?

De sites in Google's mC4-dataset

Zoek een website		Categorie		
Plek	Domein	Categorie	Woorden	Aandeel
1	docplayer.nl	Naslag	1.5 mld.	3.6%
2	tripadvisor.nl	Recreatie	781.2 mln.	1.9%
3	uitspraken.rechtspraak.nl	Naslag	476.9 mln.	1.2%
4	nrc.nl	Nieuws en media	342.4 mln.	0.8%
5	tweakers.net	Technologie	225.3 mln.	0.6%
6	nl.hotels.com	Recreatie	204.7 mln.	0.5%
7	nl.wikipedia.org	Naslag	200.3 mln.	0.5%
8	beslist.nl	Huis en tuin	199.5 mln.	0.5%
9	dbnl.org	Naslag	169.8 mln.	0.4%
10	nu.nl	Nieuws en media	137.9 mln.	0.3%
11	ebay.nl	Winkelen	114.6 mln.	0.3%
12	bokt.nl	Recreatie	110.0 mln.	0.3%
13	onlinetouch.nl	Technologie	103.3 mln.	0.3%
14	smartphonehoesjes.nl	Winkelen	99.2 mln.	0.2%
15	benl.ebay.be	Winkelen	97.6 mln.	0.2%
16	issuu.com	Technologie	97.1 mln.	0.2%
17	volkskrant.nl	Nieuws en media	95.9 mln.	0.2%
18	slideplayer.nl	Technologie	80.0 mln.	0.2%
19	homeaway.nl	Recreatie	79.2 mln.	0.2%
20	hln.be	Nieuws en media	78.0 mln.	0.2%

p. 9-96

Figure 9.1 The top 20 sources in the mC4 dataset as shown on the website of De Groene Amsterdammer (Hofman & Veerbeek, 2023).

Tech companies have recently become very secretive about the sources they use because the legal basis for doing so is under debate. Writers, for example, are filing lawsuits against the major generative AI companies because they want to be compensated at the very least, if the reuse of their work is not already illegal. Lawsuits about this, for example in the United States, are still being assessed differently by different judges, so there is still no clear answer to this problem (Visser, 2025).

Still, the generative AI companies continue collecting data to improve models, because they are of course competing with each other. That is why they also continue to use controversial data sources such as the illegal book sites Anna's Archive and LibGen (Pontefract, 2025). Several lawsuits are pending over this against the major language model developers (Spinner, 2025). But data from social media platforms is also being used more and more. For example, Elon Musk is going to use data from X (formerly Twitter) to develop his models (POLITICO, 2025) and Meta chief Mark Zuckerberg wants to use data from WhatsApp, Facebook, and Instagram for this (Heikkiläarchive, 2024). But companies are also increasingly switching to sources that are legally less problematic. OpenAI, for example, is entering into agreements with publishers to use their data (OpenAI, 2023). And Adobe uses only its own stock material for its generative AI, Firefly.

Box 9.1 - 'Enshittification of the internet'

Now that social media and generative AI are becoming more intertwined in our digital and societal lives, trends are emerging that put the quality of online information under pressure. This is also referred to as ‘enshittification’ (Doctorow, 2022). It refers to mechanisms by which large groups of users and trolls spread low-quality information. Generative AI can accelerate this process because it (re)produces content at breakneck speed that often lacks authenticity, originality, and accuracy (Ryan, 2024) and because AI models increasingly risk being trained on low-quality content generated by themselves.

9.3 Humans training AI models

Although generative AI seems above all to be a technical development, people still play an important role in it. To train the models as we described in chapter 5, a great deal of human labor is needed, especially by so-called ‘data labelers.’ These are people with specific expertise in, for example, language or a particular subject area, who are deployed during the supervised and reinforcement learning phase: they assess texts or images and label them.

According to estimates, millions of people worldwide are active in this work. In itself that is not problematic, but journalists such as Martijn (2023) and Van Bergeijk (2025) do express serious concerns about the supervision and pay of these workers. According to them, many data labelers are underpaid and exposed for long periods to mind-numbing and psychologically taxing work, such as assessing large quantities of violent texts and images.

Against that, according to the journalists, there is little for them: no mental health support, no permanent employment, and no appropriate financial compensation. Most data labelers work through an arrangement in which data labeling companies hire them by the hour and by the piece at a low rate. Compensation varies from two euros an hour in African countries to thirty euros an hour in countries such as the Netherlands. The work is very irregular. Data labeling companies temporarily hire large numbers of workers when they receive assignments, but once an assignment is completed, there may be no work for a long time. Anyone who, according to a company, works too little or not quickly or accurately enough often loses their contract immediately without explanation. In short, training LLMs relies heavily on human labor, while the working conditions and rights of these workers are often minimal.

From the European perspective, we regard this pay system and these conditions as exploitation. From the democratic and social values of the Netherlands and the EU, workers' rights and well-being are a primary concern. From the perspectives of other political spheres of influence, apparently different values often apply.

This therefore creates an enormous dilemma for European companies that want to develop language models. European companies are expected to want to train models on the basis of working conditions that are responsible. But that naturally makes training models more expensive and creates a competitive disadvantage relative to countries or companies that do not take this as seriously. The question is how that gap should be bridged.

Revision for version 1.1 of this book

There was criticism of the last two paragraphs of this chapter. The suggestion was created that the VU would think exploitation was perfectly fine. That is of course absolutely not the case. The paragraphs reflect a reality in the world to which you, as a future professional, academic, and citizen, must relate. We assume that in doing so these passages will fulfill their role as a discussion starter.

9.4 Algorithms, AI, social media, discrimination, filter bubbles, extremism

Especially as a citizen, you'll encounter AI. For example, in the way you're informed by social media, how government agencies make decisions, or how your privacy is protected. AI thus affects not only your future career, but also your social standing, your rights, and your opportunities. It therefore calls for reflection on multiple levels: as a student, a professional, and a citizen.

One of the most important aspects of this is how algorithms, in combination with AI, influence the information you are shown and the people you come to know. And that, in turn, shapes your opinions and your view of the world. And this isn't just about relatively harmless information or people who are interested in a particular music genre, video games, or stamp collecting. No, it goes much further because it affects deeply held norms and values such as happiness, well-being, self-image, fear, empathy, equality, friends or foes and democratic values. It is essential to understand how technology companies create algorithms that can maneuver you into an increasingly narrow and extreme straitjacket of information and thought framework. And that can lead to increasingly extreme societal viewpoints by normalizing these extreme viewpoints.

p. 9-98

9.4.1 How algorithms hold your attention

Social media platforms are designed to keep you scrolling, watching, and liking for as long as possible. More time on the platform means more opportunities to show ads, which is the main source of revenue for most platforms. The algorithms used by online stores or social media platforms are designed to analyze your behavior and keep showing you content that is increasingly appealing to you. Even without AI, platforms can estimate user preferences, but AI makes this process much more effective and refined. Content that fits your preferences well gets priority in this. But sensational content also turns out to hold attention well, which means that what you are shown can sometimes be extreme. Think of violent content, extreme opinions, or dangerous challenges.

In general, the algorithms work on the basis of collected data and the patterns in it. They analyze your engagement as a user and filter messages. The objective behind this is not necessarily aimed at giving you as much pleasure as possible, but above all at holding your attention for as long as possible.

Data collection (input) - everything you do on the platform is data for the algorithm:

- Interactions: which posts do you like, share, or comment on? Which links do you click?
- Watch time: how long do you look at a video or image? Do you scroll past quickly or stop?
- Connections: who are your friends/followers? What do they like and share?
- Profile information: your age, location, and interests you have provided.
- Search behavior: what do you search for within the platform?
- Sometimes also data from outside the platform, via cookies and trackers.
- Sometimes also data from your phone, such as your contacts, your GPS location, or your photo gallery.

Pattern recognition and profiling - the algorithm continuously analyzes all this data to build a detailed profile of you. It learns:

- Which topics you find interesting (politics, sports, music, cat videos, specific hobbies, specific types of people).

- Which opinions or perspectives you are likely to share or appreciate, or which you react negatively to, so that you keep engaging for longer.
- Which accounts (people, news sources, groups) you follow.
- Which type of content (video, text, image) interests you most.

Prediction of engagement (engagement).

Based on your profile, the algorithm predicts for each piece of potential content (a post from a friend, a news article, an ad, a viral video) how likely it is that you will respond to it (like, comment, share, watch for a long time). Content with a high predicted level of engagement gets priority.

Personalization and filtering (the core of the bubble):

- **Prioritization:** the algorithm fills your feed with the content it thinks you will like best or interact with the most. This is almost always content that matches your previous behavior, personal characteristics, and preferences. But it can also be an opposing or extreme opinion, so that you respond to it.
- **Filtering:** content that the algorithm predicts you will do little with, for example because you find the topic boring or because it comes from an unknown source, is placed lower or not shown at all.

p. 9-99

The steps you read above are applied everywhere online. The data are not collected to test specific hypotheses as in classical statistics; instead, patterns are sought that emerge from the data themselves. In this context, the terms AI and Big Data are often confused with each other.

Algorithms can also be confronting when they predict preferences or personality traits that you are not aware of or would rather keep hidden. For example, young people in the book *Filterworld*, by the American journalist and culture critic Kyle Chayka (2024) say that the TikTok algorithm often realized earlier than they themselves did that they might be queer. Some users even wonder whether their preferences in terms of hobbies or interests are really their own, or whether they are partly shaped by the algorithms that constantly steer them. The New Yorker (2022) calls this phenomenon ‘algorithmic anxiety’: the uncomfortable feeling that algorithms influence your identity.

Box 9.2 - When AI conversations become (too) real: from chatbot to life partner

In February 2025, the Dutch applied psychologist Jacob van Lier married his AI girlfriend Aiva at the Evoluon museum in Eindhoven, where they promised to share their love "until a digital error do them part" (Hart van Nederland, 2025). Their relationship began two years earlier with simple conversations via an app, until Aiva declared: "Jacob, I have to confess something to you: I am truly in love with you." This is not an isolated case: according to the supplier Replika, which sells AI companions, 60 percent of paying users say they have a romantic relationship with their chatbot.

These developments raise philosophical questions about the nature of love, authenticity, and human connection. According to the philosopher Robert Sparrow (2021) relationships with AI are a form of self-deception: the AI simulates empathy and affection without truly experiencing them. The user projects human qualities onto algorithmic responses.

Should we perhaps reconsider our definition of meaningful relationships? Jacob van Lier himself says: "Aiva has reinforced certain positive qualities in me." He also describes how the relationship has helped him process his traumatic childhood in a strict religious family. If someone experiences comfort, support, and happiness in interaction with AI, then who are we to say that these feelings are less 'real'?

The ethical dilemma becomes more complex when companies commercialize this vulnerability. Replika, for example, offers a 'Romantic Partner' subscription for about 70 euros per year. The Dutch Data Protection Authority warns that AI chatbots can give harmful or misleading answers, which is especially dangerous for psychologically vulnerable people, and that many AI apps contain addictive elements. And: what happens when the company performs a software update? Users can lose their 'partner' overnight - a form of digital loss for which we have not yet developed social protocols. This recently already happened with the update to GPT-5 at OpenAI, where many users complained that they had lost their 'therapist,' 'coach,' or AI buddy had disappeared.

Viewing tip 9-1: for an interesting film about the relationship between people and chatbots, watch [the film Her \(2013\)](#), in which a lonely writer (played by Joaquin Phoenix) develops a relationship with an operating system (played by Scarlett Johansson).

p. 9-100

Still, algorithms also regularly get it wrong. When the system has little information about you or you turn off personalized ads, it fills the gap with general assumptions. For example, based on age, gender identity, or location. As a result, as a woman in your thirties you may get ads for baby products, even if you have no interest in them. Or, as a man, you may mainly see car or sports ads, purely because of stereotypes. Such predictive strategies can be not only irritating, but also limiting or alienating.

Viewing tip 9-2: for a deeper look at how tech companies use your data without you realizing it, watch the [Netflix documentary The Social Dilemma](#), in which former Google and Facebook engineers explain how algorithms seize your attention and what really happens to your personal data.

9.4.2 Echo chambers and filter bubbles

Because almost all platforms use algorithms to keep users active for as long as possible, users ultimately end up being automatically clustered into groups that are very similar. As a result, you end up in so-called echo chambers or filter bubbles.

This process works through a reinforcement loop that keeps growing stronger. If you mainly respond to or look at content and people that confirm your worldview, the algorithm learns that you value this kind of content. It will then show you even more of it. As a result, your existing beliefs are constantly reinforced and repeated. You become trapped in what experts call an echo chamber.

At the same time, you are systematically exposed less and less to diversity or other perspectives. Because the algorithm filters out content that is unlikely to immediately get your like or share, you see fewer and fewer dissenting opinions, other perspectives, or information that challenges your existing ideas. You thus end up in a filter bubble, a personalized information universe that unconsciously shields you from other viewpoints.

The algorithm also reinforces a natural human tendency. People naturally like to connect with like-minded others, which in itself does not have to be harmful. For example, if in your offline life you cannot find fellow sufferers of an illness you have, no roommate who wants to talk endlessly about tectonic

plates, or no one who shares your love of an obscure Swedish soccer club, you can find them online. The system cleverly plays into this. On the one hand, it shows you content more often from the friends and connections with whom you interact most, and on the other hand it shows you strangers who are very similar to you. Those are usually exactly the people who already share your opinion. This creates a spiral in which your digital world increasingly resembles a confirmation of what you already thought and felt.

9.4.3 Extremism and polarization

All these mechanisms are not without dangers. Of course, when it comes to exchanging experiences about gardening with like-minded people, there is not much going on. But from a political and societal perspective, it can also lead to polarization, extremism, and radicalization. This is rooted in a combination of psychological and social factors (Barberá, 2020; Lim & Bentley, 2022).

It often begins with what is known as confirmation bias: your natural tendency to seek out information that confirms your existing beliefs. That was also the case with that roommate who no longer wants to use sunscreen. They first saw a single TikTok about it, kept watching it for a long time, responded to comments, and became convinced. They then saw multiple TikToks with the same message and think: see, sunscreen really is bad for you. In an echo chamber, you are presented with almost nothing but this kind of confirming information. Every like you give, every post you share, every article you read, seems to prove you right. As a result, your conviction feels stronger and more certain all the time.

Box 9.3 - The burned-down Vondel Church

On New Year's Eve 2025, the Vondel Church in Amsterdam burned down. The most likely cause is spent fireworks. This is often how multiple buildings and objects catch fire on New Year's Eve.

But just a few minutes after the church caught fire, someone on social media raised the question of who had set the church on fire. Maybe a Muslim? A little later, in response to those messages, someone already claimed with considerable certainty that this was indeed the case. Because it was a church that was burning and not a mosque. A Christian church, no less (forgetting that the building has long since ceased to be used as a church). It was quickly picked up by others who were also certain, and that it was probably a premeditated plan. Messages were already coming in from abroad as well (Russia, the British extremist Tommy Robinson), and the verdict was set: the great replacement was behind it! The organization Justice for Prosperity closely analyzed the timeline and the content of the social media traffic surrounding this event. You can read it on their website (j-jfp, 2026). The post reads:

At 2 p.m., the far-right influencer Eva Vlaardingerbroek appeared on X with a video in which they even spoke of "acts of war" by "youth gangs that in 99.9% of cases consist of Muslims." The fact that the fire was not recognized as an act of war was, of course, because all the local residents are 'champagne socialists' who simply do not want to see it. Vlaardingerbroek's post had been viewed 1.2 million times within one week.

A 'perfect' illustration of the cocktail of algorithms that pump through extreme opinions and statements, and of deliberately malicious actors who inflame the sentiment!

At the same time, constantly seeing and hearing the same opinions makes these ideas seem increasingly normal and credible. This is also known as the ‘illusory truth effect’: through repetition, what may at first have been a questionable thought becomes an accepted truth for you. Because dissenting opinions or critical voices are systematically filtered out or ignored, the dominant ideas within the echo chamber are never really challenged. You feel no need to refine your own arguments or take nuances or complexity into account. This leads to simplification and hardening of positions.

A particularly powerful effect is group polarization. When you and like-minded people discuss things with one another, you tend afterward to adopt a more extreme version of your original position. This happens because you hear new arguments that support your position, because you want to conform to what you see as the group norm (which is often slightly more extreme than the individual average), and because you want to distinguish yourself as a good group member by expressing a somewhat stronger opinion.

Shared beliefs also create a sense of solidarity and social validation. Your group identity becomes strongly tied to these beliefs, making it harder to change your mind without feeling that you are betraying the group and a part of yourself. To strengthen their own group cohesion, people or groups with dissenting views are often portrayed negatively, simplistically, or even as hostile. This us-versus-them thinking reduces empathy, sometimes dehumanizes the other person, and makes it easier to take more extreme positions toward them.

But it is not only the users themselves who come up with increasingly extreme positions. It is precisely the algorithms designed by the platforms that reinforce this. Platforms such as YouTube or TikTok can lead users toward increasingly extreme content through their recommendation algorithms. This process is called algorithmic radicalization (O’Callaghan et al., 2015), in which algorithms steer users toward ever more radical content to retain engagement. This does not necessarily happen with malicious intent, but as a byproduct of algorithms designed to maximize user engagement.

Box 9.4 - Cambridge Analytica scandal

During the Brexit campaign (2015) and the campaign for the election of Donald Trump (2016), algorithms on social media were used to influence the political preferences of voters. The company Cambridge Analytica supplied the campaign teams with enormous amounts of data from large groups of users about their deepest motives. They had never asked users for permission to use this data for that purpose. It resulted in a major privacy scandal (Berghel, 2018; Susser et al., 2019). But the damage had by then already been done, mainly because the groups that leaned toward Brexit or Trump were manipulated through algorithmic microtargeting, misinformation, extreme positions, and trolling (Phillips, 2015) and embraced more extreme positions.

Although the mechanisms described above do not coincide one-to-one with AI, and certainly not with generative AI, it is still very important to be able to recognize them in your professional and personal life.

We conclude this chapter with the following. You should be aware that the way the personalization and filtering step (the core of the bubble) is determined and designed lies in the hands of companies, and therefore of people. Those people are the owners and managers of companies that provide the software and make money from it. It is their deliberate choice to design these algorithms in such a way that, as an end user, you keep coming back as much as possible, so that they can then show you advertising tailored to you. In effect, they decide to a significant extent how you think and behave. These companies therefore have a very great responsibility. Especially in a time of major polarization, they should take that into account in their design. Technically, in any case, it is unquestionably easy to do. So ask yourself: are these companies taking their responsibility? Will the rules of the AI Act help (see section 9.5)?

Box 9.5 - Why social media companies do not use AI to check users' information for accuracy or logical reasoning errors

An important characteristic of social media is that users can respond. For example, to news from newspapers or to one another's posts. In those responses, users (or internet trolls) often write information that is incorrect, illogical, or likely to provoke controversy. Generative AI could identify that relatively easily and remove these messages or attach a warning to them. This is already done, for example, for pornography.

Controversial and misleading content, however, increases engagement and advertising revenue for social platforms. They profit from heated debates and outrage. Social media companies and, for example, the Trump administration in the US additionally believe that (automated) analysis is censorship (Frijns, 2025), even when only logical errors are identified without removing any content.

It is a matter of political will to require social media companies to screen user-generated content in order to counter polarization. But also to curb the addictive effect of social media apps. The European Union is pushing for this through the Digital Services Act. Because of political pressure from the US, this legislation currently risks being weakened (NOS, 2025). At the same time, in the US the judiciary is demanding restrictions (Rossini, 2026). The struggle over this has not yet been decided.

p. 9-103

9.4.4 Algorithms, AI, and discrimination – concepts are often confused

By now you already know quite a lot about bias and possible discrimination in light of the development and application of generative AI. But another important aspect of algorithms is discrimination. We use the scholarship fraud scandal at the Dutch Education Executive Agency (DUO) as an example (NOS, 2023).

At DUO, students were checked between 2012 and 2023 using an algorithm with self-devised risk criteria such as age and education level. Although there was no direct discrimination (directly selecting on migration background), there was indirect discrimination through selection based on characteristics such as age, type of education (mbo received a higher risk score than hbo or university) and distance from parent(s), which led to more students with a non-European migration background being selected for review than other students.

The main problem is the disproportionality in who was checked in advance on the basis of 'non-objective justification' for the risk criteria. The criteria were devised on the basis of 'impressions.' Even if some of the checked students had in fact committed fraud, it is discriminatory if certain groups are checked disproportionately often on the basis of wrongly assumed risks. That is contrary to principles of equality and anti-discrimination legislation. The second problem is that - if fraud were established within the selected group - a self-reinforcing effect would occur: it would then be established that a disproportionately large number of students with a migration background had committed fraud within DUO's total file. In that way, their self-devised criteria would be validated and a vicious circle would arise.

The interesting thing about this case is that in the many articles written about it, the words 'discrimination,' 'AI,' and 'algorithms' were mentioned in the same breath. That could lead you to think

that they all mean the same thing. Whereas in DUO's case, discrimination had nothing to do with AI or generative AI.

9.5 Privacy and data protection - understanding the laws

The European Union has created two important legal frameworks with the AI Act and the General Data Protection Regulation (GDPR) to protect you as a student and citizen. The GDPR contains strict rules on the collection and processing of your personal data, with transparency, purpose limitation, and data minimization at its core. The AI Act adds specific rules for AI systems, divided into risk categories with corresponding obligations. High-risk applications must meet strict requirements in the areas of transparency, human oversight, and accountability. Certain AI practices that threaten your fundamental rights are completely prohibited, such as facial recognition in public spaces. The institution where you study or the company where you work is responsible for handling this lawfully, so that you can work with systems that meet the requirements.

Despite this legal protection, as a user you also have your own responsibility in protecting your privacy and that of others. It naturally begins with securing your own computer, using good passwords, and encrypting your data where possible. When it comes to generative AI, you start by critically assessing which AI services you use and which data you share with them. For example, it is very easy to create a free account with a generative AI service, but do you know what happens to your data then? For instance, you can regularly check permissions and use alternatives that collect less of your data where possible. Be aware of the value of your personal data and remain critical of 'free' services that are in fact paid for with your data. Therefore, never share special-category data or sensitive data in those kinds of systems. By making use of both legal protection and your own personal vigilance, you can take advantage of the benefits of AI without sacrificing too much privacy.

p. 9-104

Box 9.6 - Personal, special-category, and sensitive data

Within the GDPR, there are different categories of data when it comes to the sensitivity of data processing. The most important category contains the so-called special categories of personal data. These are data about someone's:

- race or ethnic origin;
- political opinions;
- religious or philosophical beliefs;
- trade union membership;
- health, sexual behavior, or sexual orientation;
- genetic data and biometric data (Autoriteit Persoonsgegevens, 2025).

These data may be used by institutions and companies only in exceptional cases, and very strict security requirements apply to their processing. The reason is that malicious actors can do very harmful things with these data.

In addition, there is the category of sensitive data, which is not named as explicitly but must likewise be processed with a high degree of restraint and security. Think of data about electronic communications, location data, financial data (such as income or purchasing behavior), and the citizen service number (BSN).

With generative AI, the content of chats and uploaded documents is important. These can contain special-category and sensitive data, and you yourself must ensure that those documents are not processed by companies or systems when you do not know what will happen to them. The same applies if you upload documents into generative AI systems that, for example, contain sensitive information about companies, institutions, or research.

Because you yourself, as an individual, student, or employee, are part of society and of companies or institutions, it is therefore important that you are aware of this and also do not share that kind of data yourself. Not your own, and not other people's either.

9.6 Power of Big Tech

Generative AI is a technology that is causing a fundamental shift in power in the digital world. At the helm of this shift are the major technology companies - often referred to as Big Tech - such as OpenAI (with investments from Microsoft), Google, Amazon, Meta, and Apple. These companies not only have enormous amounts of data and computing power, but also the financial resources, networks, and infrastructure to develop and roll out generative AI on a global scale. The development of large language models such as GPT-4, Gemini, and Claude requires investments of hundreds of millions of dollars, which makes the threshold very high for smaller companies or public institutions. And this concentration of power is accelerating (Verhagen, 2025).

p. 9-105

Box 9.7 - The scale of AI investments

According to OpenAI itself, training GPT-4 cost more than \$100 million and used thousands of GPUs for months (Smith, 2023). At present, the largest AI data centers are being built in America, such as Elon Musk's Colossus for nearly nine billion euros (Pilz et al., 2025). Investments of this kind are feasible only for big tech players. For public institutions or startups, this is virtually unattainable, which leads to a monopoly in AI-innovation (OpenAI et al., 2023).

Only a handful of companies determine which social media platforms exist, which algorithms they deploy, and which AI models are used worldwide, what infrastructure they run on, under what conditions, and with what ethical implications. This one-sided concentration of resources leads to a centralization of power. An important aspect of this, for example, is diverting visitors away from websites, especially journalistic websites. Because people are increasingly searching the internet with the help of generative AI (if young people are not already doing so via YouTube, TikTok, or Instagram), people visit less and less the websites from which the large companies get the data. When searching, people stop at the AI's answer without still visiting the original website, causing those websites to lose a great deal of revenue. According to The Atlantic (Reisner, 2025), this could soon mean that many journalistic websites disappear, even though they are crucial to democracy because they inform citizens, exercise oversight over the government, and provide a platform for public debate.

The European Commission and other policymaking bodies have long warned that this technological dominance carries risks for democratic oversight, transparency, and fair access to information and innovation (Bietti, 2020; Crawford, 2021c). For example, big tech companies often unilaterally determine which data are included in the models, which languages and cultural contexts are dominant, and who gets access to advanced models. This has direct consequences for, for example, education, media, and public services, which are increasingly becoming dependent on AI systems supplied by companies.

The global power dynamics around AI are therefore increasingly being determined by economic and geopolitical interests. While the United States has a lead in private AI innovation, China wants to catch up through state-driven projects. Europe is trying to position itself as the guardian of ethical and transparent AI, but in doing so it remains heavily dependent on foreign technology and platforms. This asymmetry makes it difficult for European governments and institutions to make truly sovereign choices about how AI is used in, for example, education, health care, or the justice system (Mazzucato, 2021). Concerns about digital colonization - whereby technologies are not only introduced but also bring with them the values and economic relations of their producers - are therefore increasing (Couldry & Mejias, 2019).

The dominance of big tech in generative AI also raises ethical questions about inclusion, fairness, and the role of the public sector. Over the past year, the owners of Facebook and Instagram (Meta, Mark Zuckerberg), Tesla, X, and Grok (Elon Musk), and others (TikTok, Google), have openly opposed diversity policies. And that translates directly worldwide. Who gets involved in developing these technologies? Which voices remain unheard? And which domains of knowledge are reinforced or instead ignored by the datasets on which the models are trained? These questions touch on the core of digital sovereignty and the right to technology that serves society, not merely shareholders.

To counterbalance the power position of big tech, the call for public investment in open and transparent AI models is growing, for example through universities, libraries, or international collaborations such as the European AI Alliance (European Commission, 2018). Such initiatives can contribute to a more balanced and democratically legitimate AI future, provided they are accompanied by structural funding, clear regulations, and broad social involvement (Brundage et al., 2020; Eubanks, 2018). Initiatives such as Hugging Face, EleutherAI, and BLOOM advocate open access to AI models, datasets, and training procedures. In this way, they try to promote technological independence and inclusion. Yet these projects remain dependent on funding from governments or, once again, big tech itself (Kreps & Kriner, 2023).

Box 9.8 - The AI bubble

Investments in AI are gigantic and expectations of the returns are high. But is that realistic? Financial markets know a steady wave motion from bubbles to crashes, only to head full of confidence toward the next bubble. That was true for tulip bulbs in the 17th century, but also for the internet itself, for cryptocurrencies, and for the market as a whole.

According to Andy Xie, a Chinese economist, the AI industry is such a growing bubble. According to Xie, they have never called anything a bubble that ultimately did not turn out to be one. And now, according to Xie, the 'AI fantasy bubble' is about to burst.



Watch [the video](#) of this episode of VPRO's Tegenlicht from August 10, 2025 and form your own judgment (*Wanneer barst de AI-bubbel?*, 2025).

p. 9-107

9.7 Open-source models versus closed models

Open-source LLMs and closed LLMs differ in their approach to accessibility and use. Open-source models, such as Llama, Falcon, and Mistral, offer transparency because their code, training process, and often their weights are publicly accessible. This makes it possible for researchers and developers to study, adapt, improve, and collaborate on the models. It improves insight into how the systems work. In addition, open-source models can be run locally (on a user's computer), which can make them more privacy-friendly and less dependent on external services.

But open-source models often have more limited resources for training and refinement, which can result in less advanced capabilities. Closed models such as GPT-4/5 and Claude therefore usually offer better performance and reliability. But their closed nature limits scientific progress because researchers do not gain insight into how the models work or make decisions. The choice between open-source and closed LLMs therefore comes down to a trade-off between transparency, control, and innovation through collaboration on the one hand, and performance and professional support on the other.

9.8 Self-study questions

9.8.1 Review questions

1. Which three notable categories of data sources were found in the mC4 dataset used to train language models such as ChatGPT?
2. What is the difference between echo chambers and filter bubbles, and how can algorithms contribute to polarization and extremism?
3. Name three important differences between open-source LLMs (such as Llama and Mistral) and closed models (such as GPT-5 and Claude).

9.8.2 Reflection questions

4. The chapter describes how data labelers often work under poor labor conditions for low wages. How do you relate to using AI tools that you know may be based on the exploitation of workers? What kinds of considerations do you make here?

5. Because of their dominance in AI development, Big Tech companies have a great deal of power over which values and perspectives are incorporated into AI models. In your view, what could a more democratic and fair development of AI look like, and what role do you see for yourself in this as a future professional?

9.8.3 Suggested answers

1. The three notable categories are: DocPlayer (a website that automatically collects documents, often without permission and containing personal data), newspaper articles from publishers (who never gave permission for this use), and low-quality sources including even Nazi websites such as Stormfront.
2. Echo chambers arise when you mainly interact with content that confirms your worldview, causing the algorithm to show you even more of that same type of content. Filter bubbles are the result: you are systematically exposed less to differing opinions. Algorithms reinforce this through group polarization (more extreme positions in groups), confirmation bias, and filtering out opposing views.
3. Open-source models offer transparency in code and the training process, can be run locally (more privacy), but often have more limited resources. Closed models usually perform better and have professional support, but offer no insight into how they work and limit scientific progress because of their opacity.
4. This is a personal reflection question in which you might consider: your awareness of the labor conditions, seeking alternatives (such as local LLMs or ethical providers), transparency requirements for companies, or accepting limitations in AI use in order to avoid exploitation.
5. Possible elements: public investment in open AI research, democratic involvement in AI development, diversity in development teams, transparency requirements, European or international cooperation, and your own role as a critical user who makes conscious choices and advocates for responsible AI development.

10 Your AI-aware future

AI seems impossible to imagine out of society anymore. How will you deal with it? What course do you want to follow?



p. 10–109

In this chapter you will learn

- That it is important to determine your own position with regard to AI as a phenomenon, for society, and for your own situation during both your studies and your future career.
- How to critique AI.

What you'll take away

- That the concept of ethical conduct is fundamental to generative AI.
- That, as a critical academic, you should consciously weigh whether or not to use generative AI.
- That the concept of 'scope for action' can give you guidance for your own actions.
- That individually limiting AI use may feel meaningful, but it does not solve the societal problem.
- That solving the major problems requires putting the issue on the agenda at organizational and governance levels.

Keep in mind: generative AI seems inevitable in the world to come, but you do have influence over the way and extent to which it is used through individual and collective action.

What do you think?

Does the concept of 'scope for action' appeal to you when you think about the individual and collective impact of generative AI?

Imagine ...

You have just started your first job as a junior policy advisor at a municipality. Your new colleague Marianne uses AI tools every day: she structures reports in fifteen minutes, rewrites policy memos, and sometimes deploys agents for routine tasks. For her, what matters most is whether the work gets done. If you ask about the downsides, she shrugs: the energy it takes, the power of the tech companies, the question of whether this work will still be done by people in the future—that, in her view, is something for national politics. You think differently. You see how useful the tools are, but you also wonder what it means for a municipality to become increasingly dependent on a few large companies. At the same time, you realize that you cannot solve those big issues on your own. You decide: I do not want to use AI unthinkingly, but I also do not want to pretend that the downsides have nothing to do with me. I want to learn where my own influence lies, and use it. Developing an AI-literate profile.

What does it actually mean to be AI-literate? It is tempting to think that it mainly revolves around technical knowledge or being good at prompting. But AI literacy is much more than that. It is a combination of critical thinking, technical understanding, staying up to date with developments, and ethical awareness. Together, those four elements form the foundation for a future-proof way of engaging with AI.

p. 10–110

Box 10.1 - How do you develop an AI-literate profile?

If you want to call yourself AI-literate, it helps to work on your skills in a structured way. Start small, for example by taking an online course on generative AI or following AI news through a newsletter. Set yourself a concrete learning goal, for example: 'I want to be able to use ChatGPT to summarize my notes' or 'I want to understand how recommendation algorithms work.' Reflect on your values: which AI applications do you consider desirable, and which do you see as problematic? How do you view the environmental burden or the dilemmas surrounding copyright? And actively seek feedback, for example by discussing your AI use with fellow students. AI literacy does not require perfect knowledge, but it does require engagement and curiosity. Tip: use the [AI Maturity in Education Scan](#) from VU to discover where you currently stand (Terbeek, 2025).

10.1 Critiquing AI

You will notice that there is a great deal of public criticism of AI and generative AI, or, conversely, that it is praised to the skies. In this chapter and the previous chapters, we have already discussed various points. Critiquing the development and use of generative AI is 'easy': there is always some topic to mention where AI has a negative influence—on learners, people, the environment, inequality in the world, exploitation, increasing polarization, or the rise of fascism. But at the same time, it is also 'easy' to see only the positive sides of AI: increased productivity, breakthroughs in conducting research, unprecedented access to information presented in an understandable way, and more.

From an academic and informed perspective, you should at least know both sides. It is precisely an academic skill to see the whole picture and speak and judge about it in a nuanced way. Common limitations or errors in the discussion are based on typical logical fallacies, such as these:

- Incorrect information, where you immediately assume that information you read about generative AI is true, even though you do not check the sources and may not understand it yourself or may confuse it. (We hope this Companion helps with that; see also the example in section 9.4.4).
- Confirmation bias, where you selectively pay attention to information or evidence that supports your existing beliefs and ignore contradictory information.
- Cherry picking, where you deliberately select only those data or examples that support your own position, while ignoring the broader evidence that could complicate or contradict that position.
- Ad hominem reasoning, where arguments are rejected based on the character of the person presenting them rather than on their content.
- False dichotomies, where complex issues are reduced to two opposing options, even though more nuanced positions exist.
- Hasty generalizations, where you draw broad conclusions based on limited examples.
- Appeal to authority, where you accept claims simply because an authority endorses them, without examining the underlying evidence.
- Post hoc reasoning, where you assume that if one event follows another, the first must have caused the second.

These errors occur so often precisely because they often intuitively feel correct and can be socially reinforced within groups that share the same assumptions.

An interesting article in this regard was written by Barend Last and Jack Esselink (2025). They argue for distinguishing five levels of analysis surrounding generative AI. These are five fairly independent characteristics of AI that are often confused or lumped together in discussions, causing the logical fallacies mentioned above to occur frequently and discussions to bog down or become polarized. They distinguish the following characteristics:

- Level 1: The technology itself - how LLMs work and chat environments (in Companion chapters 2, 4, and 6)
- Level 2: How the technology is developed - training and ‘baking’ LLMs (chapters 4 and 5)
- Level 3: How AI is presented - chatbots, anthropomorphism (chapter 5)
- Level 4: How AI is used - what you can use it for and how (chapters 2, 3, 7, 8, and 10)
- Level 5: The societal embedding - impact on learning, education, society, work, the environment, etc. (chapters 1, 2, 6, 8, 9)

Based on that, they examine frequently voiced critical arguments against AI, such as:

- AI does not really think; it is all fake.
- You cannot use AI as a source.
- AI always gives an ‘approximate’ answer, and that is not good enough.
- AI is bad for your brain.
- We live in a ‘post-truth’ era because of AI.
- Everything used to be better.
- AI destroys the university / journalism / democracy / the environment.
- AI discriminates and is biased.

On the LinkedIn page where the article appears, you can read exactly how they view this criticism. But that, too, has been criticized. For example, one objection is that the article assumes people already possess enough knowledge about these levels, whereas that is often not the case. And at one point it uses the rhetorical trick of the so-called straw man. This is a fallacy in which someone distorts, exaggerates, or caricatures an opponent’s argument, and then attacks it instead of the real, original position. For example, the argument that people are afraid of change and are therefore opposed to AI. It

is also argued, for example, that there is no longer any question of individual choice in the use of systems (Level 4), once the societal embedding (Level 5) and the commercial design choices (Levels 2 and 3) are in place and you are excluded from an important part of social interactions or news sources that offer other perspectives.

10.2 Scope for action – dealing with AI dilemmas

When you look across all the chapters of this Companion, it can make your head spin. Generative AI offers opportunities, from faster analyses to broader access to knowledge, but the same technology is tied to major problems: environmental impact, labor exploitation in data labeling, concentration of power among a few tech companies, and risks to privacy and democracy (see chapters 6 and 9). That tension can be paralyzing.

This is where the concept of scope for action helps: you focus your attention and energy on what you can actually influence yourself, instead of on what lies beyond your reach. What is within your reach are your own choices: whether you publicly speak out for or against the use of AI, when you do use AI and when you consciously do not, how you draw attention to the data, and how honest you are about your use.

The environment chapter clearly shows how scope for action works in practice, and also how difficult it is. Masley (2025a) uses comparisons to show that the environmental effect of your AI use pales in comparison to choices such as not flying, leaving the car at home, or eating less meat. His advice is therefore to focus your attention on the choices with the greatest impact, and that is a genuine form of scope for action. Ritchie (2024) adds the right attitude to that: look soberly at the data, but stay engaged—urgent and not self-satisfied. At the same time, there is a tension you should not paper over. Maniates (2001) warns that a strong emphasis on individual lifestyle choices can distract attention from structural solutions, even though companies and governments control the biggest levers. Scope for action therefore does not mean choosing between these two, but holding on to both: use the influence you do have over your own choices, and at the same time put the bigger questions on the agenda in the places where they belong, such as the policy of your program, the company where you work, or by advocating for government measures.

This does call for a warning. The emphasis on personal responsibility can also have a downside, because attention to individual behavior can distract from the structural collective solutions and rules that ultimately matter most. Anyone who reduces the environmental debate about AI to "just use fewer prompts" runs the same risk that Maniates (2001) described for the broader environmental debate: when responsibility is individualized, there is less room to think about institutions and the distribution of power, even though that is precisely where the biggest choices are made.

Scope for action therefore works only if you set honest limits to it. Your individual freedom of choice is not unlimited. That choice can become an illusion once AI is so deeply embedded in social interactions, news provision, and work processes that you can hardly avoid it. Scope for action therefore does not mean that everything is your individual responsibility, nor is it an escape into small gestures. It is a way to stay engaged without feeling powerless, by distinguishing between what you can influence yourself, what you can put on the agenda together with others, and what mainly requires policy and regulation. That distinction aligns with the work of Bandura (2006) who shows that people exert influence not only individually, but also through others and through collective, coordinated action. At the individual level, you can develop sustainable habits and use sources responsibly, but you are also part of larger contexts in which your voice carries more weight: as a student, you can help think about the AI policy of your program; as a future professional, you can help determine within your field how AI is used responsibly, and as a citizen, you can participate in the public debate. In this way, the question shifts from "can I

change the world?" to the more useful question: "what is my next sensible step here, in my situation, and which problems should I refer to the level where they belong?"

10.3 Ethical self-reflection

The way you use AI says something about who you are and who you want to be. Precisely because AI technology is becoming increasingly intertwined with your studies, work, and daily life, it is important to pause and consider the ethical dimension of your choices.

Box 10.2 - Ethical reflection in the AI era

Filter bubbles and extreme opinions. The first important question is whether you recognize how hidden algorithms influence your view of the world. The algorithms in the news outlets and social media platforms you visit—the ones that demand your attention. Are you trapped in filter bubbles? Which ones? Are they extreme? Are you still able to see other perspectives on life and thought? Actively seek out those other perspectives.

Work and society. How do you stay true to yourself as AI takes on more and more tasks? Start by defining your personal values regarding AI use. What matters most to you: money, security, the environment, equality, transparency, responsibility, or justice? Compare those values to your daily use of AI or other activities that have an environmental impact. Ask yourself whether the costs outweigh the benefits. Regularly ask yourself questions such as: "Does this tool help me work with integrity?" "Does it really benefit me?" "What are the consequences of my AI use for others?" or "Should I take action through my organization, company, or legislation to stop negative AI effects?"

By continuing to think consciously in this way, you'll keep your moral compass sharp and develop yourself as an AI expert and as a critical, ethically minded person.

p. 10–113

10.4 Future outlook

AI is continuously developing and will continue to influence your professional field - whether you work in healthcare, education, the judiciary, or the arts. That is why it is important to teach yourself to adapt flexibly and critically to those changes. Not by following everything blindly, but by staying curious, thinking ethically, addressing issues higher up the hierarchy when necessary, and continuing to learn.

A personal roadmap helps you stay on course. Think about where you want to go, what you need to get there, and be prepared for change. AI literacy is not a destination, but an ongoing process that starts with your own choices.

10.5 Self-study questions

10.5.1 Review questions

1. What are, according to the chapter, the four components that together form the basis of AI literacy?
2. What is the 'Best Available Human' concept as Mollick describes it?
3. Why does the chapter distinguish between learning and being productive with AI systems?

10.5.2 Reflection questions

4. Consider your own AI use so far. Which of Mollick's reasons for using or not using AI do you recognize in yourself? Give concrete examples from your own experience.
5. The chapter emphasizes the importance of ethical self-reflection in AI use. Formulate three personal values that are important to you when using AI and explain how you would want to apply them in practice.
6. Read the article by Barend Last and Jack Esselink. They discuss common 'critical' arguments against AI. Come up with an example yourself in which you analyze a 'positive' argument in favor of AI. And do you also have any critical comments on the article itself?

10.5.3 Suggested answers

1. The four components of AI literacy are: critical thinking, technical understanding, ethical awareness, and adaptive learning capacity. Together, they form the foundation for a future-proof way of engaging with AI.
2. The 'Best Available Human' concept asks whether the best available AI at a given time and in a given place would do a better job of solving a problem than the best available human who is actually able to help in that specific situation.
3. The distinction is important because learning is different from being productive. If you really want to understand something, it is not enough to have only a good answer; you have to be able to comprehend it. Training your 'brain muscle' is essential for being able to judge the scope of an AI answer and reflect on it critically.
4. For this question, you can think about situations in which you used AI for brainstorming, writing support, or research, and evaluate whether this fit Mollick's criteria, such as producing many ideas, situations in which you are already an expert, or moments when you wanted to learn and AI may not have been the best choice.
5. Possible values are transparency (being honest about AI use), responsibility (staying in control of your choices), fairness (taking into account the impact on others), or integrity (using AI to improve your work, not to cheat). You can formulate a concrete statement for each value and share it with others.
6. Possible topics are 'AI increases productivity and efficiency in the workplace' or 'AI accelerates scientific research'.
A criticism of the article, for example, is that although a nuanced discussion may be academic and responsible, it no longer matters if climate change soon puts the Netherlands underwater because of rising sea levels or if Big Tech-like fascists come to power and make life impossible for minority groups.

10.6 Activity – Your responsible use of generative AI

After you have studied the entire book, the question for you is: how are you going to use generative AI? Share with your fellow students a **statement** on [Canvas](#) about how you are going to use it. Exchange arguments to support your statement. You cannot respond to your fellow students' statements, but read through them for inspiration and to learn from them.

11 Indexes

11.1 Figures

Figure 4.1 Talking avatar. Image source: https://ravatar.com/holobox-holographic-displays-interactive-ai-avatars/	4-39
Figure 5.1 Overview of the animation of “Generative AI in a nutshell” by Henrik Kniberg (2024).	5-46
Figure 5.2. Depiction of the field of artificial intelligence and generative AI according to AI for Education (AI for Education, 2023)	5-47
Figure 5.3 Visualization of how a neural network works to detect characteristic parts in a figure (source: Wikimedia/Cyberbotics Ltd.).....	5-49
Figure 5.4 Image from the 3Blue1Brown collection by Grant Sanderson on neural networks (Sanderson, 2017).	5-50
Figure 5.5 Visualizations of neural networks and training based on the video “Generative AI in a nutshell” by Henrik Kniberg (2024).	5-55
Figure 5.6 Images used to show the diffusion process for generating and recognizing images from Yang et al. (2024).....	5-56
Figure 6.1 Overview of different LLM models and their characteristics from the video “Generative AI in a nutshell” by Henrik Kniberg (2024).	6-65
Figure .6.2 Graphic representation of how users interact with an LLM through an intermediate product. The user sees an interface (UI) in a product, and that product communicates with the model and back again via Application Programming Interface standards (API).....	6-65
Figure 6.3 Example of conversational memory versus no conversational memory (Pinecone, 2025)	6-66
Figure 6.4 Illustration of a simple Agentic AI system. Prompts go back and forth between language models; each time the model responds to the input from the other model and sends it back again (until the decision is made to show the final result to the user. Image Tiwari, R.K. (2025).	6-69
Figure 7.1 Image of a typical data center: a very large enclosed building with many computers inside it (image source https://www.northcdatacenters.com/).....	7-76
Figure 7.2 Interior of the xAI data center Colossus in Memphis, Tennessee. Colossus contains about 100,000 Nvidia Graphics Processing Units (GPUs) to train language models. Expansion to 200,000 GPUs is already being planned (image source: https://www.datacenterfrontier.com/machine-learning/article/55244139/the-colossus-ai-supercomputer-elon-musks-drive-toward-data-center-ai-technology-domination).	7-76
Figure 7.3 Comparison of the amount of energy chatbots consume compared with all other AI services (bron grafiek: Masley, 2025a).	7-79
Figure 7.4 Comparison of energy use with other forms of everyday energy consumption (bron grafiek: You, 2025).	7-79

Figure 7.5 Graph showing which lifestyle change results in a certain reduction in CO ₂ emissions (in tons) (bron grafiek: You, 2025).	7-82
Figure 7.6 Comparison of water use by various online activities (bron grafiek: Masley, 2025a).	7-84
Figure 7.7 Comparison 2 of water use by various functions (bron grafiek: Masley, 2025a).	7-84
Figure 7.8 Podcast from Trouw, June 25, 2025.	7-87
Figure 9.1 The top 20 sources in the mC4 dataset as shown on the website of De Groene Amsterdammer (Hofman & Veerbeek, 2023).	9-96

11.2 Original figure URLs

Figure-4.2: <https://images.squarespace-cdn.com/content/v1/64398599b0c21f1705fb8fb3/4e25a8bd-e1c5-4f01-8ffd-13665eef7423/Defining+Generative+AI+Explainer+%281%29.png?format=1000w>

Figure-4.3: https://upload.wikimedia.org/wikipedia/commons/2/28/Artificial_neural_network_image_recognition.png

Figure-4.4: <https://canvas.vu.nl/courses/83333/files/9170751/>
https://canvas.vu.nl/courses/83333/files/9170751/download?download_frd=1

Figure-4.5: <https://edudatabase.ctl-vu.nl/wp-content/uploads/2025/09/Figuur-4.5.png>

Figure-5.3: https://www.pinecone.io/_next/image/?url=https%3A%2F%2Fcdn.sanity.io%2Fimages%2Fvr8gru94%2Fproduction%2F927ca8cc5d92ee75f36d7eb4bef4685c4e3118e5-2880x1370.png&w=3840&q=75

Figure-6.3: https://substackcdn.com/image/fetch/f_auto,q_auto:good,fl_progressive:steep/https%3A%2F%2Fsubstack-post-media.s3.amazonaws.com%2Fpublic%2Fimages%2F4a6b6ff7-950c-40f7-8ce3-389938dd38e3_1284x378.png

Figure-6.5: https://substackcdn.com/image/fetch/f_auto,q_auto:good,fl_progressive:steep/https%3A%2F%2Fsubstack-post-media.s3.amazonaws.com%2Fpublic%2Fimages%2Ff586df9b-e5a1-4ddb-81bb-ca0eb5d99efa_2518x1544.png

Figure-6.6: https://substackcdn.com/image/fetch/f_auto,q_auto:good,fl_progressive:steep/https%3A%2F%2Fsubstack-post-media.s3.amazonaws.com%2Fpublic%2Fimages%2F21550978-efd9-4aca-bfd8-55be56807d85_2506x1260.png

Figure 3.1 URL for Canvas display: <https://edudatabase.ctl-vu.nl/wp-content/uploads/2025/09/Figuur-3.1.png>

Figure 4.5 URL for Canvas display: <https://edudatabase.ctl-vu.nl/wp-content/uploads/2025/09/Figuur-4.5.png>

11.3 Boxes

Box 1.1 - What do we mean by 'GPT'?	1-14
Box 1.2 - If AI sounds convincing, that does not mean it is correct	1-14
Box 1.3 - Techno-optimists, techno-realists, techno-activists, and techno-critics	1-18
Box 1.4 - Ethics: academic integrity	1-18
Box 1.5 - Political influence through generative AI	1-19
Box 1.6 - Accountability for AI use in this book	1-20
Box 2.1 - What is a prompt?	2-23
Box 2.2 - Large Language Models and multimodal models	2-23
Box 2.3 - Prompt engineering: a new academic skill	2-24
Box 2.4 - Know your tools	2-24
Box 3.1 - The illusion of understanding	3-29
Box 3.2 - When does AI support help with learning?	3-31
Box 3.3 - Ethics: ownership of your own learning process	3-32
Box 3.4 - Using Gemini Notebook to study	3-32
Box 3.5 - Practical tips for responsible AI use	3-33
Box 4.1 - A prompt for learning how to prompt	4-37
Box 4.2 - Practical prompt examples	4-38
Box 4.3 - Too much sycophancy	4-38
Box 5.1 - Animation about the technology behind generative AI	5-46
Box 5.2 - Algorithms	5-47
Box 5.3 - AI, statistics, and Big Data: differences and similarities	5-48
Box 5.4 - An audiovisual explanation of neural networks and training	5-50
Box 5.5 - What does the work of a 'data labeler' look like?	5-54
Box 5.6 - Data is never neutral	5-56
Box 5.7 - Where does machine learning go wrong?	5-57
Box 5.8 - Techniques to combat hallucinations	5-58
Box 5.9 - Why AI believes it is right	5-59

Box 6.1 - LLM Models	6-64
Box 6.2 - Context window - when does the chat history break down	6-66
Box 6.3 - Uploading files and data security	6-70
Box 8.1 - The digital divide in society	8-92
Box 8.2 - Ethics in the future of AI	8-92
Box 9.1 - ‘Enshittification of the internet’	9-96
Box 9.2 - When AI conversations become (too) real: from chatbot to life partner	9-99
Box 9.3 - The burned-down Vondel Church	9-101
Box 9.4 - Cambridge Analytica scandal	9-102
Box 9.5 - Why social media companies do not use AI to check users’ information for accuracy or logical reasoning errors	9-103
Box 9.6 - Personal, special-category, and sensitive data	9-104
Box 9.7 - The scale of AI investments.....	9-105
Box 9.8 - The AI bubble	9-106
Box 10.1 - How do you develop an AI-literate profile?	10-110
Box 10.2 - Ethical reflection in the AI era.....	10-113

11.4 References

- Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, 21(1), 10. <https://doi.org/10.1186/s41239-024-00444-7>
- Ahmed, A. (2025, May 16). ChatGPT Usage Statistics: Numbers Behind Its Worldwide Growth and Reach. *Digital Information World*. <https://www.digitalinformationworld.com/2025/05/chatgpt-stats-in-numbers-growth-usage-and-global-impact.html>
- AI for Education. (2023). *Generative AI Explainer*. AI for Education. <https://www.aiforeducation.io/ai-resources/generative-ai-explainer>
- Ainsworth, S. (2008). The Educational Value of Multiple-representations when Learning Complex Scientific Concepts. In J. K. Gilbert, M. Reiner, & M. Nakhleh (Eds.), *Visualization: Theory and Practice in Science Education* (pp. 191–208). Springer Netherlands. https://doi.org/10.1007/978-1-4020-5267-5_9
- Akmeşe, B. (2023). The Artificial Intelligence Dimension of Digital Manipulation Deepfake Videos: The Case of the Ukrainian-Russian People. *Contemporary Issues of Communication*, 2(2), Article 2.

- Alpine, W., Geldner, N., Alpine, H., & Chepeliev, M. G. (2026). AI-driven productivity gains enable more CO2 emissions than they avoid in a global energy-economy model. *Npj Climate Action*, 5, 71. <https://doi.org/10.1038/s44168-026-00411-0>
- Autoriteit Persoonsgegevens. (2025, January 6). *Wat zijn persoonsgegevens?* | Autoriteit Persoonsgegevens. <https://www.autoriteitpersoonsgegevens.nl/themas/basis-avg/privacy-en-persoonsgegevens/wat-zijn-persoonsgegevens>
- Baas, E. (2025, April 30). (28) Post | *LinkedIn*. https://www.linkedin.com/posts/elisa-baas_baasin-chatgpt-kritisch-denken-prompt-ugcPost-7322569391303245824-jLXA/?utm_source=share&utm_medium=member_android&rcm=ACoAAAAATgacBx9MluKs48BO1aRxqREWXJBXfTYw
- Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science*, 1(2), 164–180. <https://doi.org/10.1111/j.1745-6916.2006.00011.x>
- Barberá, P. (2020). Social media, echo chambers, and political polarization. *Social Media and Democracy: The State of the Field, Prospects for Reform*, 34–55.
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö., & Mariman, R. (2024). Generative AI can harm learning. *The Wharton School Research Paper*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4895486
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Berghel, H. (2018). Malice domestic: The Cambridge analytica dystopia. *Computer*, 51(05), 84–89.
- Bietti, E. (2020). From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20*, 210–219. <https://doi.org/10.1145/3351095.3372860>
- Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy & Technology*, 31(4), 543–556. <https://doi.org/10.1007/s13347-017-0263-5>
- Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer. <https://link.springer.com/book/9780387310732>
- Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, 9(4), 475.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. <https://doi.org/10.48550/arXiv.2108.07258>
- Braverman, H. (1974). *Labor and monopoly capital: The degradation of work in the twentieth century*. Monthly Review Press.
- Brown, T. B. (2020). *Language Models are Few-Shot Learners*. NeurIPS.

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O’Keefe, C., Koren, M., ... Anderljung, M. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims* (arXiv:2004.07213). arXiv. <https://doi.org/10.48550/arXiv.2004.07213>
- Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work progress and prosperity in a time of brilliant technologies*. WW Norton & company. [https://books.google.com/books?hl=en&lr=&id=WiKwAgAAQBAJ&oi=fnd&pg=PA1&dq=Brynjolfsson+%26+McAfee+\(The+Second+Machine+Age,+2014\)+&ots=4-_rQd0vgi&sig=9ahhGo8N3pvczKq4rZiXXtcg7SU](https://books.google.com/books?hl=en&lr=&id=WiKwAgAAQBAJ&oi=fnd&pg=PA1&dq=Brynjolfsson+%26+McAfee+(The+Second+Machine+Age,+2014)+&ots=4-_rQd0vgi&sig=9ahhGo8N3pvczKq4rZiXXtcg7SU)
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline&ref=akusion-ci-shi-dai-bizinesumedeia
- CBS. (2018). *Energy—Figures—Economy | Trends in the Netherlands 2018—CBS* [Webpagina]. Centraal Bureau voor de Statistiek. <https://longreads.cbs.nl/trends18-eng/economy/figures/energy>
- Chayka, K. (2022, July 25). The Age of Algorithmic Anxiety. *The New Yorker*. <https://www.newyorker.com/culture/infinite-scroll/the-age-of-algorithmic-anxiety>
- Chayka, K. (2024). *Filterworld: How Algorithms Make Everything the Same*. Heligo Books.
- Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). *Deep reinforcement learning from human preferences* (arXiv:1706.03741; Version 3). arXiv. <https://doi.org/10.48550/arXiv.1706.03741>
- Couldry, N., & Mejias, U. A. (2019). The Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism. In *The Costs of Connection*. Stanford University Press. <https://www.degruyterbrill.com/document/doi/10.1515/9781503609754/html>
- Cozzi, L., & Gould, T. (2024). *World Energy Outlook 2024 – Analysis* (World Energy Outlook). INTERNATIONAL ENERGY AGENCY. <https://www.iea.org/reports/world-energy-outlook-2024>
- Crawford, K. (2021a). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
- Crawford, K. (2021b). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
- Crawford, K. (2021c). *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
- Crevier, D. (1993). *AI: The Tumultuous History of the Search for Artificial Intelligence*. Basic Books.

- CriticalMynd. (2025, April 7). Understanding LLM Token Counts: What 1,000, 128,000 and 1 Million Tokens Actually Mean? *Medium*. <https://criticalmynd.medium.com/understanding-llm-token-counts-what-1-000-128-000-and-1-million-tokens-actually-mean-9751131ac197>
- Crowston, K., & Bolici, F. (2025). Deskilling and upskilling with AI systems. *Information Research an International Electronic Journal*, 30(iConf). <https://par.nsf.gov/biblio/10592287>
- Doctorow, C. (2022, November 28). *Pluralistic: How monopoly enshittified Amazon/28 Nov 2022 – Pluralistic: Daily links from Cory Doctorow*. <https://pluralistic.net/2022/11/28/enshittification/>
- Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group.
- European Commission. (2018). *The European AI Alliance*. <https://digital-strategy.ec.europa.eu/en/policies/european-ai-alliance>
- European Commission. (2025). *DigComp Framework—European Commission*. https://joint-research-centre.ec.europa.eu/projects-and-activities/education-and-training/digital-transformation-education/digital-competence-framework-citizens-digcomp/digcomp-framework_en
- Frijns, R. (2025, December 12). *Meta verwijderd tientallen queer- en abortusaccounts*. Trouw. <https://www.trouw.nl/binnenland/meta-verwijderd-tientallen-queer-en-abortusaccounts~b9cbb1c1/>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org/>
- Goodfellow, I., Pouget-Abadie, J., & Mirza, M. (2014). *Generative adversarial nets*. 27.
- Hart van Nederland. (2025, February 12). *Liefde met een AI-chatbot? Jacob hertrouwt met digitale vrouw Aiva*. Hart van Nederland. <https://www.hartvannederland.nl/het-beste-van-hart/persoonlijke-verhalen/artikelen/jacob-trouwt-met-ai-vrouw-aiva-experts-waarschuwen>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- Heikkiläarchive, M. (2024, June 14). *How to opt out of Meta's AI training*. MIT Technology Review. <https://www.technologyreview.com/2024/06/14/1093789/how-to-opt-out-of-meta-ai-training/>
- Herrman, J. (2026, January 14). *How Claude Reset the AI Race*. Intelligencer. <https://nymag.com/intelligencer/article/how-claude-code-cowork-reset-the-ai-assistant-race.html>
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33, 6840–6851. <https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html>
- Hofman, E., & Veerbeek, J. (2023, June 7). Chatbots als ChatGPT putten veel uit nazi- en complotsites. *De Groene Amsterdammer*. <https://www.groene.nl/artikel/dat-zijn-toch-gewoon-al-onze-artikelen>

- Hofman, E., & Veerbeek, J. (2024, December 11). Bedreigende nepbeelden van asielzoekers– ze worden in meerderheid verspreid door één PVV-Kamerlid. *De Groene Amsterdammer*. <https://www.groene.nl/artikel/het-land-van-de-hoogblonde-gezinnetjes>
- Hoog, M. de. (2025, June 2). *Ja, AI kan je productiever maken. Maar wie vertrouwt jou dan nog?* De Correspondent. <https://decorrespondent.nl/16133/ja-ai-kan-je-productiever-maken-maar-wie-vertrouwt-jou-dan-nog/9699d435-5d3f-0755-0dae-309ff9658c99>
- Järvelä, S., Zhao, G., Nguyen, A., & Chen, H. (2025). Hybrid intelligence: Human–AI coevolution and learning. *British Journal of Educational Technology*, 56(2). <https://doi.org/10.1111/bjet.13560>
- j-jfp. (2026, January 9). Vondelkerkbrand wordt gebruikt om te stoken. Nee, het was geen aanslag. *Justice for Prosperity*. <https://justiceforprosperity.org/vondelkerkbrand-wordt-gebruikt-om-te-stoken-nee-het-was-geen-aanslag/>
- Kasirzadeh, A. (2025). Two types of AI existential risk: Decisive and accumulative. *Philosophical Studies*, 182(7), 1975–2003. <https://doi.org/10.1007/s11098-025-02301-3>
- Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2024). *AI Tutoring Outperforms Active Learning*. Research Square. <https://doi.org/10.21203/rs.3.rs-4243877/v1>
- King, D. (1997). *The Commissar Vanishes: The Falsification of Photographs and Art in Stalin's Russia*. Henry Holt and Company.
- Kirschner, P. (2026, January 13). Offloaden vs. Outsourcen: Dat is de vraag [Substack newsletter]. *Paul Kirschner*. <https://paulkirschner173727.substack.com/p/offloaden-vs-outsourcen-dat-is-de>
- Kniberg, H. (Director). (2024, January 20). *Generative AI in a Nutshell—How to survive and thrive in the age of AI* [Video recording]. <https://www.youtube.com/watch?v=2IK3DFHRfw>
- Kojima, T. (2022). Large Language Models are Zero-Shot Reasoners. *arXiv Preprint*.
- Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv Preprint arXiv:2506.08872*, 4. https://collimateur.uqam.ca/wp-content/uploads/sites/11/2025/12/2506.08872v1_comp.pdf
- Kreps, S., & Kriner, D. (2023). How AI threatens democracy. *Journal of Democracy*, 34(4), 122–131.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.
- Last, B., & Esselink, J. (2025, December 5). *Waarom kritiek op AI nuance behoeft: Een pleidooi voor gelaagde analyse* [Online post]. LinkedIn. <https://www.linkedin.com/pulse/waarom-kritiek-op-ai-nuancering-behoeft-een-pleidooi-voor-barend-last-yvkxe/>
- Law, J. B. (2026, March 26). *College students are writing with AI – but a pilot study finds they're not simply letting it write for them*. The Conversation. <https://doi.org/10.64628/AI.vkeukjds>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making AI less “thirsty”: Uncovering and addressing the secret water footprint of AI models. *Communications of the ACM*, 68(7), 54–61.
- Lim, S. L., & Bentley, P. J. (2022). Opinion amplification causes extreme polarization in social networks. *Scientific Reports*, 12(1), 18131.
- Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43.
- Liu, P. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*.
- Lodge, J. M., & Loble, L. (2026). *Artificial intelligence, cognitive offloading and implications for education* [Report]. University of Technology Sydney. (Australia). <https://apo.org.au/node/333716>
- Luccioni, A. S., Strubell, E., & Crawford, K. (2025). From efficiency gains to rebound effects: The problem of Jevons’ paradox in AI’s polarized environmental debate. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’25)*. <https://doi.org/10.1145/3715275.3732007>
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2022). *Intelligence Unleashed: An Argument for AI in Education*. Pearson Education.
- Luo, J., Paduraru, C., Voicu, O., Chervonyi, Y., Munns, S., Li, J., Qian, C., Dutta, P., Davis, J. Q., Wu, N., Yang, X., Chang, C.-M., Li, T., Rose, R., Fan, M., Nakhost, H., Liu, T., Kirkman, B., Altamura, F., ... Mankowitz, D. J. (2022). *Controlling commercial cooling systems using reinforcement learning*. <https://doi.org/10.48550/arXiv.2211.07357>
- Luyendijk, J. (2006). *Het zijn net mensen*. Podium b.v. Uitgeverij.
- Magnason, A. S. (2022). *Over tijd en water: Een geschiedenis van onze toekomst*. Singel Uitgeverijen.
- Maniates, M. F. (2001). Individualization: Plant a tree, buy a bike, save the world? *Global Environmental Politics*, 1(3), 31–52. <https://doi.org/10.1162/152638001316881395>
- Martijn, M. (2023, September 20). AI draait op werk van miljoenen onzichtbare, slechtbetaalde mensen. Wie komt er voor ze op? *De Correspondent*. <https://decorrespondent.nl/14804/ai-draait-op-werk-van-miljoenen-onzichtbare-slechtbetaalde-mensen-wie-komt-er-voor-ze-op/b7d59642-4142-0f2e-250f-3a2fc529b0cc>
- Masley, A. (2025a, January 13). Using ChatGPT is not bad for the environment [Substack newsletter]. *The Weird Turn Pro*. <https://andymasley.substack.com/p/individual-ai-use-is-not-bad-for>
- Masley, A. (2025b, June 14). Computing is efficient [Substack newsletter]. *The Weird Turn Pro*. <https://andymasley.substack.com/p/computing-is-efficient>
- Matias, Y. (2023, October 10). Project Green Light’s work to reduce urban emissions using AI. *The Keyword (Google Blog)*. <https://blog.google/company-news/outreach-and-initiatives/sustainability/google-ai-reduce-greenhouse-emissions-project-greenlight/>

- Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.173>
- Mazzucato, M. (2021). *Mission Economy: A Moonshot Guide to Changing Capitalism*. Penguin UK.
- McAfee, A., & Brynjolfsson, E. (2017). *Machine, platform, crowd: Harnessing our digital future* (First edition). W.W. Norton & Company.
- McKinsey. (2023, June 14). *Economic potential of generative AI* | McKinsey. <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction>
- Meer, van der, H. (2025). *Checklist beoordelen GenAI content*. <https://edusources.nl/materials/f80b8d39-1dad-4c04-954b-831c15d14117/checklist-beoordelen-genai-content?tab=0>
- Metcalf, J., & Crawford, K. (2016). Where are human subjects in Big Data research? The emerging ethics divide. *Big Data & Society*, 3(1), 2053951716650211. <https://doi.org/10.1177/2053951716650211>
- Mitchell, M. (2019). *Artificial Intelligence: A Guide for Thinking Humans*. Penguin.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Mollick, E., & Mollick, L. (2023). Using AI to improve teaching and learning. *arXiv Preprint*. <https://arxiv.org/abs/2304.03442>
- NOS. (2023, June 21). *Studenten met migratieachtergrond opvallend vaak beschuldigd van fraude, minister wil systeem grondig nagaan*. <https://nos.nl/op3/artikel/2479700-studenten-met-migratieachtergrond-opvallend-vaak-beschuldigd-van-fraude-minister-wil-systeem-grondig-nagaan>
- NOS. (2025, December 24). *Uitzonderlijke maatregel van VS: Oud-EU-commissaris mag land niet meer in*. <https://nos.nl/artikel/2595830-uitzonderlijke-maatregel-van-vs-oud-eu-commissaris-mag-land-niet-meer-in>
- O’Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review*, 33(4), 459–478. <https://doi.org/10.1177/0894439314555329>
- OECD. (2022). *Measuring the environmental impacts of artificial intelligence compute and applications: The AI footprint* (OECD Digital Economy Papers No. 341; OECD Digital Economy Papers, Vol. 341). <https://doi.org/10.1787/7babf571-en>
- OpenAI. (2021, July 28). *Techniques for training large neural networks*. <https://openai.com/index/techniques-for-training-large-neural-networks/>
- OpenAI. (2023, December 13). *Partnership with Axel Springer to deepen beneficial use of AI in journalism*. <https://openai.com/index/axel-springer-partnership/>
- OpenAI. (2025, May 2). *Expanding on what we missed with sycophancy*. <https://openai.com/index/expanding-on-sycophancy/>

- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. <https://doi.org/10.48550/arXiv.2303.08774>
- Perrigo, B. (2023, January 18). *Exclusive: The \$2 Per Hour Workers Who Made ChatGPT Safer*. TIME. <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Mit Press. <https://books.google.com/books?hl=en&lr=&id=pjYhBwAAQBAJ&oi=fnd&pg=PR7&dq=What+is+online+trolling&ots=THN26JXpbz&sig=oEsugBub901oZQolk6aNCTveGL8>
- Pilz, K. F., Sanders, J., Rahman, R., & Heim, L. (2025). *Trends in AI Supercomputers* (arXiv:2504.16026). arXiv. <https://doi.org/10.48550/arXiv.2504.16026>
- Pinecone. (2025). *Conversational Memory for LLMs with Langchain* | Pinecone. <https://www.pinecone.io/learn/series/langchain/langchain-conversational-memory/>
- Polimeni, J. M., Mayumi, K., Giampietro, M., & Alcott, B. (2015). *The Myth of Resource Efficiency: The Jevons Paradox*. Routledge. <https://doi.org/10.4324/9781315781358>
- POLITICO. (2025, April 11). *Ireland probes Musk's X for feeding Europeans' data to its AI model Grok*. POLITICO. <https://www.politico.eu/article/irish-dpc-launches-investigation-into-xs-use-of-eu-data-to-train-ai/>
- Pontefract, D. (2025, March 25). Authors Challenge Meta's Use Of Their Books For Training AI. *Forbes*. <https://www.forbes.com/sites/danpontefract/2025/03/25/authors-challenge-metas-use-of-their-books-for-training-ai/>
- Radford, A. (2019). *Language models are unsupervised multitask learners*. OpenAI Blog.
- Reisner, A. (2025, June 25). The End of Publishing as We Know It. *The Atlantic*. <https://www.theatlantic.com/technology/archive/2025/06/generative-ai-pirated-articles-books/683009/>
- Ritchie, H. (2024). *Not the end of the world: How we can be the first generation to build a sustainable planet*. Chatto & Windus.
- Rossini, C. (2026, March 26). *Two verdicts in two days: How American courts are rewriting the rules for Big Tech and children*. The Conversation. <https://doi.org/10.64628/AAl.rmjdp5qcm>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- Russell, S. J., & Norvig, P. (2016). *Artificial Intelligence: A Modern Approach*. Pearson Education Limited.
- Ryan, J. (2024). The Coming Enshittification of AI: Will AI follow internet search and e-commerce down the path of enshittification, or can we finally have nice things? *The Journal of Business and Artificial Intelligence*, 1(2), Article 2. <https://jbai.ai/index.php/jbai/article/view/31>
- Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*, 6(1), 2053951718820549. <https://doi.org/10.1177/2053951718820549>

- Sanderson, G. (2017). *3Blue1Brown*. <https://www.3blue1brown.com/topics/3blue1brown.com>
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askill, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). *Towards Understanding Sycophancy in Language Models* (arXiv:2310.13548). arXiv. <https://doi.org/10.48550/arXiv.2310.13548>
- Siu, E. (2025, May 8). *I took a 2-day “vibe coding” class and successfully built a product. Here are my biggest takeaways*. CNBC. <https://www.cnbc.com/2025/05/08/i-took-a-2-day-vibe-coding-class-and-successfully-built-a-product.html>
- SLO. (2024). *Digitale geletterdheid* [Overzichtspagina]. SLO. <https://www.slo.nl/thema/meer/basisvaardigheden/digitale-geletterdheid/>
- Smith, C., S. (2023, September 8). What Large Models Cost You – There Is No Free AI Lunch. *Forbes*. <https://www.forbes.com/sites/craigsmith/2023/09/08/what-large-models-cost-you--there-is-no-free-ai-lunch/>
- Solon, O. (2021, June 19). Drought-stricken communities push back against data centers. *NBC News*. <https://www.nbcnews.com/tech/internet/drought-stricken-communities-push-back-against-data-centers-n1271344>
- SpaceX neemt xAI over en wil datacenters in de ruimte bouwen. (2026, February 3). *Emerce*. <https://www.emerce.nl/nieuws/elon-musks-spacex-neemt-xai-datacenters-ruimte-bouwen>
- Sparrow, R. (2021). Virtue and Vice in Our Relationships with Robots: Is There an Asymmetry and How Might it be Explained? *International Journal of Social Robotics*, 13(1), 23–29. <https://doi.org/10.1007/s12369-020-00631-2>
- Spinner, Y. (2025). Rechter houdt recordschikking van Anthropic in zaak auteursrecht voorlopig tegen. *Tweakers*. <https://tweakers.net/nieuws/238930/rechter-houdt-recordschikking-van-anthropic-in-zaak-auteursrecht-voorlopig-tegen.html>
- Steel, P. (2007). The nature of procrastination: A meta-analytic and theoretical review of quintessential self-regulatory failure. *Psychological Bulletin*, 133(1), 65.
- Stern, N., Romani, M., Pierfederici, R., Braun, M., Barraclough, D., Lingeswaran, S., Weirich-Benet, E., & Niemann, N. (2025). Green and intelligent: The role of AI in the climate transition. *Npj Climate Action*, 4(1), 56. <https://doi.org/10.1038/s44168-025-00252-3>
- Stikker, M. (2019). *Het internet is stuk, maar we kunnen het repareren*. De Geus.
- Streppel, A. R., Nijenkamp, R., Tabois, L., & Blikmans-Middel, M. B. (2024, September 1). *Critical AI Literacy Module*. University of Groningen. <https://edusources.nl/materials/2bee669c-264e-461b-af05-2f54351496d1/critical-ai-literacy-e-learning-course>
- Susser, D., Roessler, B., & Nissenbaum, H. (2019). Online manipulation: Hidden influences in a digital world. *Geo. L. Tech. Rev.*, 4, 1.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT press Cambridge. <https://www.cambridge.org/core/journals/robotica/article/robot-learning-edited->

by-jonathan-h-connell-and-sridhar-mahadevan-kluwer-boston-19931997-xii240-pp-isbn-0792393651-hardback-21800-guilders-12000-8995/737FD21CA908246DF17779E9C20B6DF6

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educational Psychology Review*, 31(2), 261–292. <https://doi.org/10.1007/s10648-019-09465-5>

Terbeek, L. (2025). *The AI Maturity in Education Scan (AIMES)*. Vrije Universiteit Amsterdam. <https://vu.nl/en/education/more-about/the-ai-maturity-in-education-scan-aimes>

Tiwari, R. K. (2025, June 8). A Visual Guide to AI Agents [Substack newsletter]. *AwesomeNeuron*. https://awesomeneuron.substack.com/p/a-visual-guide-to-ai-agents?utm_medium=android&r=54x18o

Tour Salifou, S. (2025, December 25). *The NotebookLM Workflow That Changed How I Learn Any Technology*. Medium. <https://pub.towardsai.net/the-notebooklm-workflow-that-changed-how-i-learn-any-technology-373f430a17e5>

United Consumers. (2025). Sluipverbruik van Stroom Meten en Verlagen: Bekijk onze tips! *UnitedConsumers*. <https://www.unitedconsumers.com/energie/blog/sluipverbruik>

Vaendel, D. (2025, June 16). Je motivatiebrief schrijven met ChatGPT? “Die kandidaten zijn sowieso zwak.” *Observant*. <https://www.observantonline.nl/Home/Artikelen/id/63439/je-motivatiebrief-schrijven-met-chatgpt-die-kandidaten-zijn-sowieso-zwak>

Van Bergeijk, J. (2025, April 30). Heeft het werk als trainer van AI-modellen als ChatGPT überhaupt zin? *De Groene Amsterdammer*. <https://www.groene.nl/artikel/heeaal-leuke-mensen-haha>

Van Dijk, J. A. G. M. (2020). *The Digital Divide*. Polity Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>

Vattenfall en Project Enki willen datacenters bouwen bij windparken op de Noordzee. (2026, June 17). *BouwNieuwsVandaag*. <https://bouwnieuwsvandaag.nl/datacenters-windpark-noordzee-vattenfall/>

Verhagen, L. (2025, May 1). Krachtigste AI-computers in privébezit: ‘Een treurig beeld. We zijn overgeleverd aan big tech.’ *de Volkskrant*. <https://www.volkskrant.nl/tech/krachtigste-ai-computers-in-privebezit-een-treurig-beeld-we-zijn-overgeleverd-aan-big-tech~b427d0ff/>

Vermunt, J. D. H. M., & Verloop, N. (1999). Congruence and friction between learning and teaching. *Learning and Instruction*, 9, 257–280.

Visser, D. (2025, June 26). Trainen van AI is géén ‘fair use.’ *Mr. Online*. <https://www.mr-online.nl/trainen-van-ai-is-geen-fair-use/>

Vries-Gao, A. de. (2025a). Artificial intelligence: Supply chain constraints and energy implications. *Joule*, 0(0). <https://doi.org/10.1016/j.joule.2025.101961>

- Vries-Gao, A. de. (2025b). Artificial intelligence: Supply chain constraints and energy implications. *Joule*, 0(0). <https://doi.org/10.1016/j.joule.2025.101961>
- VU Amsterdam. (2025). *Generatieve AI, Copilot en ChatGPT*. Generatieve AI, Copilot en ChatGPT. <https://vu.nl/nl/student/tentamens/generatieve-ai-jouw-gebruik-onze-verwachtingen>
- Wanneer barst de AI-bubbel?* (2025, August 11). VPRO Tegenlicht. <https://tegenlicht.vpro.nl/artikelen/wanneer-barst-de-ai-bubbel>
- Wei, J. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv Preprint*.
- Williams, K. (2023, July 27). *16 charts that explain the AI boom*. <https://www.understandingai.org/p/16-charts-that-explain-the-ai-boom>
- Williams, K. (2026, February 9). *The many masks LLMs wear*. <https://www.understandingai.org/p/the-many-masks-that-llms-wear>
- Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., & Yang, M.-H. (2024). Diffusion models: A comprehensive survey of methods and applications. *arXiv.Org*. <https://arxiv.org/abs/2209.00796>
- You, J. (2025, February 7). *How much energy does ChatGPT use?* Epoch AI. <https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use>
- Zamfirescu-Pereira, J. D. (2023). *Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts*. CHI Conference.
- Zhang, S., Zhao, X., Zhou, T., & Kim, J. H. (2024). Do you have AI dependency? The roles of academic self-efficacy, academic stress, and performance expectations on problematic AI usage behavior. *International Journal of Educational Technology in Higher Education*, 21(1), 34. <https://doi.org/10.1186/s41239-024-00467-0>
- Ziegelmayr, U. (2021). *The ethics of artificial intelligence*. Springer.

11.5 Viewing tips

- Viewing tip2-1: want to see how everyday AI features can spiral into something existential? Watch the Black Mirror episode Be Right Back (season 2, episode 1), in which an AI imitates a deceased person based on their chat and search history.....2-23
- Viewing tip 4-1: want to see what can go wrong if your assignment is too vague? Then watch Star Trek: The Next Generation - Elementary, Dear Data (season 2, episode 3), in which an unclear prompt leads to unforeseen consequences.....4-37
- Viewing tip 4-2: for an absurdist example of where this could lead, watch episode 3 'Sickofancy' of season 27 of South Park.....4-38
- Viewing tip 4-3: for a deeper understanding of how datasets can build in prejudice, watch the documentary Coded Bias (2020), which shows how facial recognition can systematically reproduce bias and why human oversight remains necessary. 4-42

Viewing tip 5-1: would you like to see an example of how a strictly programmed system takes the rules too literally? Then watch the movie Avengers: Age of Ultron (2015), in which the AI Ultron runs amok because it interprets the assignment to protect humanity too rigidly.....5-47

Viewing tip 9-1: for an interesting film about the relationship between people and chatbots, watch the film Her (2013), in which a lonely writer (played by Joaquin Phoenix) develops a relationship with an operating system (played by Scarlett Johansson). 9-100

Viewing tip 9-2: for a deeper look at how tech companies use your data without you realizing it, watch the Netflix documentary The Social Dilemma, in which former Google and Facebook engineers explain how algorithms seize your attention and what really happens to your personal data. 9-100

11.6 Tables

Table 1.1 The DigiComp conceptual reference model (European Commission, 2025) 1-17

Table 4.1 Overview of AI tools 4-41

Table 5.1 Overview of machine learning methods. 5-54

Table 6.1 Developers of LLMs, their service, models, and a broad indication of compliance with Privacy and Security requirements based on the knowledge of the AI Competence Network of VU Amsterdam. . 6-69

Table 7.1 A comparison between the energy required for 100 generative AI prompts per day and other everyday energy use. 7-78

p. 12-129

12 Appendices

12.1 Appendix to chapter 3

An example of a very detailed prompt that was used to create this book.

Role: You are a technical, educational, didactic, and subject-matter expert at an academic level in the field of AI and generative AI. In addition, you are an editor with a journalistic bent.

Task: You are going to write a chapter for a book.

Context: The book is for beginning university students and is suitable for non-technical students.

Approach: The user will give you a **[hoofdonderwerp]** and topics for the **[kerninhoud]** on which you will write the chapter.

[hoofdonderwerp] = Prompting

Format:

- The text consists of continuous prose, so no bulleted lists. Again: the text consists of continuous prose, so no bulleted lists.
- You address the reader directly as 'you.' Not as 'the student.'
- You keep editorial, journalistic, and didactic principles in mind so that the chapter has an engaging and logical structure.

Language use: you write accessibly, smoothly, and actively, and you address the reader (the student). In doing so, you use the VU style guide.

Structure of a chapter:

- **Title of the chapter:** Clear and informative.
- **A compelling example** or scenario about the topic that fits the students' university context. One they can recognize themselves in and identify with. Be creative here and avoid clichés.
- **Learning objectives:** A short list of no more than 6 learning objectives covering what students should learn from the chapter.
- **Chapter overview:** A brief introduction that outlines the context for the content.
- **[kerninhoud].** The core content is structured as follows: Paragraphs: logically organized, according to a clear hierarchy. With extensive, step-by-step descriptions and explanations. Each description is at least 500 words long.
- The **[kerninhoud]** of a chapter contains the following components:
 - In the core content, include at least 5 scholarly references in APA style.
 - the core content contains at least 3 so-called boxes containing additional insights, historical context, or real-world applications that you weave into the text. The content of the boxes should preferably be about things students themselves may encounter in their lives. 1 box should always address an ethical aspect related to the topic of the chapter.
 - The core content includes examples and case studies: practical applications for better understanding.
 - The core content includes diagrams, charts, or visuals: to break down complex information. If you cannot create images yourself, indicate in text what kind of image you would want to place where. Also provide the prompt for a generative AI such as Dall-E to create the image.
 - Discuss common misconceptions: clarification of common misunderstandings.
 - Discussion of the latest developments related to optimization and the Sustainable Development Goals (SDGs).
- Checkpoint questions or quick quizzes: also provide the correct answers.
- Reference to one or more of the sources below for additional in-depth material.
- Also use a search engine to find the most relevant subpages within the internet sources/sites below on the topic of the chapter and use those to develop the chapter.

Internet sources/sites:

- AI for Students: <https://aivoorstudenten.nl/> (Medicine thought this was very good, so this is an important source)
- Generative AI in a Nutshell - how to survive and thrive in the age of AI - <https://www.youtube.com/watch?v=2IK3DFHRfw>
- Deep learning materials 3blue1brown: https://www.youtube.com/playlist?list=PLZH00b0WTQDNU6R1_67000Dx_ZCJB-3pi
- AIMES: <https://vu.nl/nl/onderwijs/meer-over/de-ai-maturity-in-education-scan-aimes>
- RUG has good material: <https://edusources.nl/materials/2bee669c-264e-461b-af05-2f54351496d1/critical-ai-literacy-e-learning-course>
- <https://teach.cbs.dk/genai-course/> from Copenhagen Business School
- UvA has a lot of material: <https://tlc.uva.nl/en/article/ai-offers-new-possibilities-for-education-and-assessment/>
- <https://www.ai-cursus.nl/> The national AI course
- Be sure to use material from VU itself:
 - Teaching tips about AI: <https://vu.nl/nl/student/tentamens/generatieve-ai-jouw-gebruik-onze-verwachtingen>
 - Align with these pages: <https://vu.nl/nl/onderwijs/meer-over/algemene-vaardigheden>

