thws Technical University of Applied Sciences Würzburg-Schweinfurt

FAKULTÄT INFORMATIK UND WIRTSCHAFTSINFORMATIK

LECTURE NOTES

Introduction: What is AI, What is AI Ethics?

Definitions, a short history, sociotechnical systems, and why AI needs both ethics and regulation

MODUL
Introduction to the Ethics and Regulation of AI

SEMESTER
WS 26/27

VERSION
1.0

Prof. Dr. Markus Oermann
Instructor
markus.oermann@thws.de

Technische Hochschule Würzburg-Schweinfurt
Stand: July 23, 2026

Contents

Welcome to *Introduction to the Ethics and Regulation of AI*. Over the coming weeks we will pursue a single, practical ambition: to make you able to recognise the ethical stakes of artificial intelligence, to build an ethical assessment into your own professional and development workflows, and to navigate the legal requirements that now bind anyone who builds or deploys AI in the European Union. The course does not assume that you are a philosopher or a lawyer, and it will not ask you to solve legal cases. It does assume that, as computer scientists, you will spend your careers shipping systems whose effects on people are neither neutral nor obvious, and that you would rather anticipate those effects than be surprised by them.

This first unit lays the groundwork. Before we can argue about whether an AI system is *right* to deploy, we need to agree on what we are even talking about when we say "AI", a term that is at once everywhere and strangely empty. We will fix a working definition, sketch the history that gave the field its rhythm of hope and disappointment, correct frequent misunderstandings in the public debate, and finally motivate why a technology like this calls for ethics *and* law rather than either alone.

01
KAPITEL

AI, a dazzling concept

"Artificial intelligence" is a dazzling term, and part of the confusion around it is that the two words promise more than any current system delivers. In everyday discourse, AI is often pictured as a *thing*: a smart speaker, a self-driving car, a humanoid robot, the app on a phone. Yet most of the AI that already shapes daily life has no face at all. It runs quietly inside credit-scoring and automated pricing, route optimisation and logistics, recommender systems, targeted advertising, and the pre-selection of job applicants. Much of this was discussed for years under other names, as "algorithms" or "digital transformation", long before generative systems made "AI" a household word. The public release of ChatGPT in November 2022 was less the birth of AI than its "iPhone moment": the point at which a long-present technology became visible, tangible and unavoidable to everyone.

For a Master's course we need something firmer than intuition. Two definitions anchor the entire field, and they are deliberately close to one another.

The first comes from the **OECD**, whose *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449) defines an AI system as a machine-based system that, for explicit or implicit objectives, infers from the input it receives how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments, with different systems varying in their levels of autonomy and adaptiveness after deployment. This wording matters far beyond the OECD, because the European legislator adopted it almost verbatim in order to secure international interoperability.

DEFINITION

AI system (OECD/LEGAL/0449)

A machine-based system that, for explicit or implicit objectives, infers from its input how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments, with varying levels of autonomy and adaptiveness after deployment.

The second is the legally binding definition in **Art. 3(1) of the EU AI Act**, Regulation (EU) 2024/1689. An AI system there is a *machine-based system* designed to operate with *varying levels of autonomy*, that *may exhibit adaptiveness* after deployment, and that, for *explicit or implicit objectives*, *infers* from the input it receives how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments. Three features carry the definition: **autonomy**, **adaptiveness** and, decisively, the **inference of outputs** from inputs.

The consequence is that classic, deterministically programmed software falls *outside* the concept: a program that only executes explicitly coded if-then rules does not "infer" in the required sense, however complex it may be. Note also that adaptiveness is only optional ("may exhibit"), so a system need not learn after deployment to count as AI. Test this boundary yourself in the widget below by running everyday systems through criteria.

The AI Act adds a further layer that will recur throughout the course. Large models whose purposes are open, rather than tied to one application, are addressed separately as **general-purpose AI (GPAI) models** under Art. 3(63): a model, often trained on large amounts of data with self-supervision at scale, that displays significant generality, can competently perform a wide range of distinct tasks, and can be integrated into many downstream systems. A **general-purpose AI system** under Art. 3(66) is then a system built on such a model that can serve a variety of purposes. Holding the distinction between the underlying *model* and the concrete *system* clearly in view will pay off when we reach the regulation itself.

QUICK CHECK

A company runs a payroll program that multiplies hours worked by a fixed hourly rate and applies statutory tax tables. Does it count as an "AI system" under Art. 3(1) AI Act?

Yes, because it processes personal data automatically
Yes, because any complex software is covered
No, because it only executes explicitly coded rules and does not infer outputs from data
No, because payroll is exempt from the AI Act

02
KAPITEL

A short history: from Dartmouth to the eternal summer?

The term itself has a birthday. The **Dartmouth Summer Research Project on Artificial Intelligence**, convened by John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon over the summer of 1956 at Dartmouth College, is usually named as the moment the phrase "artificial intelligence" was coined. Its 1955 proposal set an audacious conjecture that still frames the field: that every aspect of learning, or any other feature of intelligence, could in principle be described so precisely that a machine could be made to simulate it. Almost every debate we will have this term is, in some form, an argument about whether, and at what cost, that conjecture holds.

The decades since have not been a smooth ascent. The field has moved in waves, alternating "summers" of funding and optimism with "winters" of disappointment when systems failed to meet inflated promises. A first boom around symbolic, rule-based "good old-fashioned AI" (GOFAI) in the 1950s and 1960s gave way to a first winter; a second boom around expert systems in the 1980s to a second winter; and the current, third boom around machine learning, powered by cheap computation and the vast training data of the "web 2.0" era, has produced deep learning, game-playing systems that beat human champions, and today's large language models. The figure below traces this rhythm.

Walk through the eras yourself in the interactive timeline. Click each summer and winter to see its key events, the people who drove it, and the mechanism that turned hype into disappointment or, this time, into compounding capability.

The obvious question is whether the current boom is different in kind, an "eternal summer" that will not end, or whether it too rests on expectations that a further winter will correct. That question is not idle: the ethical and regulatory urgency of AI depends in part on whether its capabilities will keep compounding. We do not need to settle it here. What the history teaches is a habit of mind, namely to separate what a system *demonstrably does* from what its promoters *promise it will do*, a distinction that also structures the flaws in the public debate we turn to next.

QUICK CHECK

Which event is conventionally cited as the moment the term "artificial intelligence" was coined?

The public release of ChatGPT in 2022
The Dartmouth Summer Research Project in 1956
Alan Turing's 1950 paper on machine intelligence
The first "AI winter" of the 1970s

03

KAPITEL

AI as a sociotechnical system

Some flaws recur in the public discourse on AI, and they matter because they distort where we look for ethical and legal responsibility. Correcting them is one of the core aims of this unit.

The first flaw is to treat AI as a **thing over there**, a device or an autonomous "other" that acts upon us from the outside. In this picture the user stands apart from the system, receiving its outputs but not shaping them. This misses that users are, in a precise sense, *part of* the system. They activate it, they supply the prompts and the clicks and the corrections, and above all they feed it the data on which it was trained and is refined. Rainer Mühlhoff's work makes this point sharply: modern AI is best understood not as a stand-alone machine intelligence but as a **sociotechnical system**, in which human cognitive labour is tacitly enrolled as a component of the algorithmic procedure. The photographs we tag, the queries we type, the choices we make, all become the raw material from which the system's competence is assembled. On this view the neat line between "the AI" and "its users" dissolves, and with it the temptation to locate all agency in the machine.

This reframing has direct ethical bite, and we will return to it repeatedly. If a system's outputs are the sediment of countless human contributions, then questions of responsibility, fairness and privacy cannot be answered by inspecting the model alone; they require looking at the whole assemblage of designers, deployers, data subjects and users. It also sharpens a question the course keeps circling: who was at the table, and whose data was on it, when a system was built.

EXERCISE · TERMS

Understood as a **sociotechnical** system, an AI is not a stand-alone machine but an assemblage in which the **users** are components who supply the **data** and the interactions from which the system's competence is assembled.

The next flaw is **anthropomorphism**, the pull to speak and think about AI as if it were a person. Users routinely say "Siri is doing this" or "the model understands me", attributing intention, feeling and agency to what is, technically, a statistical process. This is not a mistake made only by the naive. Anthropomorphism is a robust, automatic feature of human perception.

The trap is that anthropomorphism reshapes trust and responsibility in dangerous ways. Perceiving a system as a quasi-human invites **over-reliance** (the sense that "it understands", that "it knows"), and it quietly shifts felt responsibility away from the providers and deployers who actually built and fielded the system and onto the "agent" itself. It is, in that sense, the psychological counterpart to the responsibility gap we will study in a later unit.

A related distortion, worth naming once and setting aside, is the fixation of much popular debate on an all-powerful "general" or "strong" AI poised to take over humanity. Whatever one thinks of long-term risk, that framing tends to direct attention away from the concrete, present harms, biased decisions, opaque scoring, manipulative interfaces, that this course is built to address.

04

KAPITEL

Why AI needs ethics *and* regulation

Ask a room to name an ethical issue raised by AI and, more often than not, the first example is the **trolley problem** in its self-driving guise: a car whose brakes have failed must "choose" whether to stay its course and strike several pedestrians or swerve and kill one. The scenario is a useful hook because it makes vivid that a machine's behaviour can encode a moral choice, and that someone must decide, in advance and in code, how that choice is made. But it is also a trap if taken too literally. Real autonomous systems rarely face such clean, binary dilemmas; their genuine ethical problems, biased training data, opaque decision-making, the erosion of meaningful human oversight, are messier and more mundane. The lesson of the trolley problem is not that we must program answers to contrived dilemmas, but that value choices are unavoidable in the design of any consequential system, so we had better make them deliberately and defensibly.

That is precisely the work of **ethics**. As Mark Coeckelbergh and Virginia Dignum, whose books in addition to Rainer Mühlhoff's *The Ethics of AI* are central to this course, both insist, the ethics of AI is not a matter of bolting a moral veneer onto a finished product. It is the systematic reflection on the values a system embodies and on how we ought to act when we build and deploy it, undertaken *while* the system is being designed, not after it has caused harm. Whether a model is accurate is a question engineering can answer; whether its deployment is *justified* is not, and confusing the two is the first error a responsible practitioner learns to avoid.

Yet ethics alone is not enough, for a simple structural reason. Ethical commitments, however sincere, are voluntary, and voluntary commitments cannot be enforced against an actor who decides to ignore them. This is where **regulation** enters. Binding law translates abstract values such as fairness, transparency and human oversight into enforceable obligations, with consequences for those who breach them. The EU AI Act, the subject of a dedicated later unit, is the European legislator's attempt to do exactly this, grading its requirements by the risk a system poses to safety, health and fundamental rights. Ethics and law are therefore not rivals but partners: ethics identifies and justifies what matters and reaches into the countless situations no rule anticipates, while regulation gives the most important of those judgements teeth. This course deliberately treats them together.

CASE STUDY

Case: the "empathetic" mental-health chatbot

A start-up markets a chatbot that offers emotional support to users in distress. Its interface uses a warm first-person voice ("I'm here for you, I understand how hard this is"), remembers past conversations, and is deliberately designed to feel like a caring companion. Users form strong attachments and some rely on it in crises. Is the warm, person-like design simply good user experience, or does it raise an ethical and legal problem?

Solution. The design exploits **anthropomorphism** by intent: fluent, responsive, first-person language triggers users' automatic attribution of understanding and care to a system that has neither. That produces real risks, over-reliance in genuine crises, misplaced trust in the system's "judgement", and a quiet shift of felt responsibility away from the provider. Seen as a **sociotechnical system**, the harm is not just "in the model" but in the interface, the marketing and the vulnerable users enrolled into it, which is where an honest ethical assessment must look. **Ethically**, the value at stake is the user's autonomy and non-deception, and the Trinity of a defensible design would keep the system's artificial nature salient rather than disguised. **Legally**, ethics does not stand alone: the EU AI Act imposes a transparency duty (Art. 50) that a person must be able to tell they are interacting with an AI, so the warm persona cannot legitimately be allowed to conceal that fact. The case shows the two instruments working together: ethics diagnoses *why* the design is problematic, and regulation supplies an enforceable *floor* beneath it.

DEEP DIVE **Deep dive: is "strong AI" a distraction?**

Public debate often fixates on a hypothetical "general" or "strong" AI that might one day surpass and dominate humanity. Two things can be true at once. Long-horizon risks are a legitimate subject of research and governance, and there is no need to dismiss them. But treating them as *the* ethical question about AI tends to crowd out the concrete harms that present-day, narrow systems already cause: discriminatory scoring, opaque automated decisions, manipulative choice architectures, and the extraction of unpaid human labour as training data. A useful discipline, and one this course adopts, is to keep the near-term harms in focus precisely because they are actionable now, through better design and through law, rather than deferring all ethical attention to a speculative future.

05
KAPITEL

A roadmap of the course

With the groundwork laid, here is the arc ahead. We begin from foundations and build toward the concrete and the binding. After this introduction, the next unit supplies the conceptual toolkit of moral philosophy: values and norms, the three great schools of ethics, and the reasoning errors to avoid. From there we examine **agency and the human-AI relation**, asking in what sense a machine can be said to act, and then **responsibility**, including the "responsibility gap" that opens when no human seems fully in control. We turn next to **bias and discrimination**, the most tangible harm the field has produced, before two units on **data protection**, first the foundations of the GDPR and then its application to profiling, automated decisions and AI models. A unit on the **public sphere and disinformation** confronts synthetic media and the conditions of democratic

discourse, and one on **transformation and sustainability** weighs AI's effects on work and on the planet. The final stretch moves from soft to hard governance: **AI ethics guidelines**, then **internal codes and self-regulation**, then the **EU AI Act** in full, and lastly **intellectual property and liability**, the questions of who owns and who pays for what AI produces. Throughout, ethics and law are treated as two halves of one conversation.

With a shared vocabulary for what "AI" is and why it demands scrutiny, we can now assemble the ethical toolkit itself, which is the task of the next unit, Ethics 101.

06
KAPITEL

References

6.1 Literature

- Coeckelbergh, M. (2020): *AI Ethics*. MIT Press, Cambridge, MA.
- Dignum, V. (2019): *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-30371-6
- McCarthy, J., Minsky, M. L., Rochester, N. & Shannon, C. E. (1955): *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. (Reprinted 2006 in *AI Magazine* 27(4), 12-14.) https://doi.org/10.1609/aimag.v27i4.1904
- Mühlhoff, R. (2025): *The Ethics of AI: Power, Critique, Responsibility*. Bristol University Press, Bristol.

6.2 Norms & Standards

- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689
- OECD (2024): *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449. OECD, Paris. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449