thws Technical University of Applied Sciences Würzburg-Schweinfurt

FAKULTÄT INFORMATIK UND WIRTSCHAFTSINFORMATIK

LECTURE NOTES

Agency & the Human-AI Relation

Moral status, machine ethics, sociotechnical entanglement and the nudge

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

In the previous unit we assembled the toolkit of moral philosophy, from virtue ethics through utilitarianism to Kant's deontology, and asked whether these traditions still have something to tell us about AI. This unit turns that question on the systems themselves: when an AI acts, decides and is spoken of as an actor, what exactly is its standing in a moral constellation? We first clarify three foundational concepts, moral status, moral patient and moral agent, then confront the machine-ethics claim that machines could be moral agents, examine why humans so readily read intention into moving shapes, follow Actor-Network Theory in redistributing agency across humans and things, and finally see how the very design of AI interfaces steers behaviour through nudges. The through-line is a single problem: to whom, or to what, do we attribute agency in the digital age?

01
KAPITEL

Three concepts: status, patient, agent

1.1 What it means to have moral standing

Three terms bring order to almost every ethical debate about a new kind of entity.

Moral status is the property of being morally considerable at all, of being something that can be treated rightly or wrongly.

A **moral patient** is an entity with moral status that is *affected* by a morally relevant action; it can be wronged, harmed or benefited.

A **moral agent**, by contrast, is an entity that can *act* with reference to right and wrong, that can make moral judgements and, crucially, bear responsibility for what it does.

The decisive insight is that these two roles come apart. Every moral agent is also a patient, but not every patient is an agent. An infant or a sentient animal has moral status and can plainly be wronged, yet we do not hold it responsible for its conduct. Patiency and agency are two distinct axes, not one scale, and much of the confusion about AI dissolves once they are kept separate.

DEFINITION

Moral patient vs. moral agent

A moral patient *undergoes* morally relevant actions and can be wronged; a moral agent *performs* them, can judge right from wrong, and bears responsibility for the outcome.

Applied to AI, the two questions split cleanly. *Can an AI be a moral patient?* Some scholars argue for a "technological person" endowed with rights of its own, but legislators have so far declined to adopt any such "electronic personhood". *Can an AI be a moral agent?* That is the claim of the school of machine ethics, examined below. And *is AI merely a tool*, developed and deployed by humans for their ends and to be treated as such? The honest answer, developed across this

unit, is that AI has qualities that exceed those of a simple tool, yet still falls short of full moral agency, which is why the central human moral agents remain in view and the question of their responsibility carries over into the next unit.

EXERCISE · TERMS

An entity that can be wronged but cannot answer for its own conduct, such as an infant, is a moral **patient** but not a moral **agent**. Granting AI its own rights would require recognising it as a moral **patient**, a step the legislator has so far refused.

02

KAPITEL

Machine ethics: can machines be moral agents?

2.1 Accountability without responsibility

The strongest version of the agency claim comes from **machine ethics**, the programme of building moral reasoning into the technology itself rather than leaving it to the humans around it. Its most influential philosophical defence is Floridi and Sanders' argument from *levels of abstraction*. Observed at the right level, they contend, an artificial system can be interactive, autonomous and adaptive enough to count as an agent, and where its actions can be good or bad it is a *moral* agent. The move that makes this palatable is a careful distinction: such an artificial agent can be morally **accountable** for its effects without being morally **responsible** for them. Responsibility, on their account, presupposes mental states, intention and a capacity to answer for oneself that machines lack; accountability requires only that the source of a morally loaded action can be identified and, if necessary, corrected or removed. AI can therefore be the *source* of moral good and harm while the *responsibility* rests elsewhere.

DEFINITION

Accountable vs. responsible (Floridi & Sanders 2004)

An artificial agent can be *accountable*, identifiable as the source of a morally loaded action, without being *responsible*, which presupposes the consciousness and intention machines do not possess.

This is why we cannot simply demand that an AI act "voluntarily" or "in good conscience". Voluntariness and awareness are exactly the properties it does not have, so the machine-ethics agent is at best a truncated agent: it can be the origin of morally weighty behaviour, but the buck

does not stop with it. Keeping accountability and responsibility apart is what lets us take AI's agency seriously without pretending it is a person, and it sets up the responsibility gap that the next unit dissects.

QUICK CHECK

On Floridi and Sanders' account, in what sense can an artificial agent be "moral"?
It has genuine intentions and can be blamed like a person.
It can be accountable as the source of good or harmful effects, without being responsible in the full sense.
It becomes a moral patient with its own rights.
It only counts as moral once it is conscious.

03
KAPITEL

Why we see agents everywhere: the intentional stance

3.1 Heider and Simmel's moving shapes

First, watch the video. Continue reading afterwards!

What did you see here? Probably a chase, children playing, or a family argument. The fascinating thing about it is that, in reality, only geometric shapes were moving here. Long before there were chatbots, psychology had already shown how eagerly humans attribute agency to things that have none. In a classic 1944 experiment, Fritz Heider and Marianne Simmel showed viewers a short animation of a large triangle, a small triangle and a circle moving around a rectangle with an opening. Asked what they saw, almost no one described geometric shapes in motion. They described a bully chasing a couple, a hero rescuing a victim, a jealous quarrel: full stories of intention, fear and desire projected onto three abstract forms. The finding is robust and cross-cultural, and it names a deep human disposition to read minds into motion.

Philosophers call the resulting habit the **intentional stance**: we make sense of a system's behaviour by treating it *as if* it had beliefs, desires and goals, because doing so is a fast and usually effective predictive strategy. It shapes our everyday language about technology, in which "Siri is doing this" and "the algorithm wants you to stay" come naturally, and it explains why conversational AI feels like a partner rather than a program. For an ethics of AI the disposition cuts both ways. It is what makes AI socially usable, and it is also what produces over-trust and automation bias, the tendency to defer to a system precisely because it presents as a knowing agent. Anthropomorphism is thus not a harmless quirk; it quietly shifts how much authority we grant a machine.

EXERCISE · TERMS

Heider and Simmel (1944) demonstrated that people spontaneously attribute intentions and feelings to moving **geometric** shapes. The habit of explaining a system's behaviour as if it had beliefs and goals is called the **intentional** stance.

CASE STUDY

Case: the grief chatbot

A start-up offers a "memorial" chatbot trained on a deceased person's messages, so that the bereaved can keep "talking" to them. Users report feeling comforted, but some describe deep distress when the bot says something the real person never would have said, and a few struggle to stop using it. Where does the moral weight sit?
Solution. The bot is not a moral agent: it has no intentions and cannot mean what it says, and following Floridi and Sanders it is at most *accountable* as the source of its outputs, never *responsible* for the harm they cause. What is doing the moral work is the user's intentional stance, powered by exactly the disposition Heider and Simmel documented: the bereaved person reads a mind, and a beloved one, into a statistical text generator. The responsibility therefore lies with the human moral agents who designed and deployed a system engineered to be anthropomorphised in a setting of acute vulnerability. This is a foretaste of the next unit: the harm is real, the machine cannot answer for it, and the burden falls on the designers who foresaw, or should have foreseen, the over-attribution their product invites.

04
KAPITEL

Redistributing agency: Actor-Network Theory

4.1 Agency as an effect, not a possession

Since the 1980s, the social studies of technology have pressed a more radical thought: perhaps agency was never the exclusive property of conscious subjects to begin with. **Actor-Network Theory (ANT)**, associated above all with Bruno Latour, attributes agency not only to people and their organisations but to *all* active entities, human and non-human alike. On this view agency is not an intention that a subject holds but an **effect** that a network produces. Humans and things act together in networks oriented toward shared goals, and social phenomena are understood as the product of that network-forming. ANT thereby tries to dissolve the modern Western split between the social and the technical, between culture and nature, treating them as woven together rather than opposed.

Latour's favourite illustration is deliberately mundane. A university asks all staff to keep an open-door policy, but Professor X wants quiet and installs a door-closer on his office door. The actor-network "door-closer plus professor" now jointly enacts a closed-door policy: neither the human intention alone nor the device alone produces the outcome, but their combination does. The door-closer is not a neutral instrument that merely executes a human will; it *delegates* and *materialises* a program of action, and it keeps enforcing it long after the professor has stopped thinking about it. Agency, in short, is distributed across the assemblage, with no single node holding all of it.

DEFINITION

Actor-Network Theory (Latour)

Agency is not the intention of a conscious subject but an *effect* produced by networks of human and non-human actants acting together toward shared goals.

The pay-off for AI is considerable. Modern systems increasingly perform tasks with social effects comparable to human actions: assessing credit and insurance risk, steering cars, ships and planes, recommending content, drafting documents, composing music. Looking at such a system through the ANT lens, we stop asking the unanswerable question "does the AI have a mind?" and ask instead a tractable one: which human and non-human actants are enrolled in this network, and what program of action does their combination enforce? That reframing is exactly what makes the responsibility analysis of the next unit possible, because it exposes the many hands, and many things, through which an outcome is jointly produced.

QUICK CHECK

What is the core move Actor-Network Theory makes about agency?

It proves that AI systems are conscious.
It restricts agency to humans and their organisations.
It treats agency as an effect of human and non-human networks, not as the intention of a lone subject.
It denies that technology has any social effect.

05
KAPITEL

AI is not just a tool

5.1 Affordances, sociotechnical systems and human computation

If ANT loosens the grip of the tool model from the side of philosophy, economics and the social studies of technology loosen it from the side of fact. Complex digital technologies are not just means to given ends; assessed as **sociotechnical systems**, they are platforms and infrastructures with powerful microeconomic effects, network effects and lock-in, that carry macroeconomic weight. They also carry **affordances**, features that invite and shape particular uses, and through them they mould cultural practices and even users' sense of themselves. Providers do not merely serve pre-existing ends; they partly *create* the ends their applications satisfy, and they stabilise a "new normal" that their business model needs in order to be accepted.

A vivid demonstration is **human computation**, popularised by Luis von Ahn (cf. von Ahn/Dabbish 2004). His "purposeful games" harvested human cognitive labour for useful output: in the ESP Game two strangers independently labelled the same image, and agreement produced a reliable tag, generating high-quality training data for image search while the players believed they were merely playing. The same logic drives **reCAPTCHA**, where solving a puzzle to prove you are human simultaneously digitises books or labels street scenes for machine-learning systems. From the web-2.0 era onward this became a general strategy: services are offered not for direct profit but for the *data*, later monetised through personalised advertising. Users supply the labour and the raw material that make AI possible, usually without noticing, which is precisely why the human-AI relation cannot be reduced to a user wielding a tool.

DEFINITION

Human computation

A design strategy that harnesses human cognitive work, often through games or CAPTCHAs, to produce data or labour that machines cannot yet generate on their own.

06
KAPITEL

The nudge: designing the choice

6.1 Choice architecture and libertarian paternalism

The clearest place where the design of a system becomes a lever on human behaviour is the **nudge**. Coined by Thaler and Sunstein (2008), a nudge is a deliberate feature of the **choice architecture**, the way options are presented, that predictably steers behaviour without forbidding any option or changing economic incentives. Choice architecture is unavoidable: options must always be arranged *somehow*, through spatial layout on an interface, the number of alternatives offered, the way each is described, and above all the **default**. Placing the salad bar at the start of the canteen line, switching organ donation from opt-in to opt-out, or reminding a taxpayer that "most of your neighbours have already filed" all leave every choice formally open while making one of them far more likely.

The mechanism is psychological. As the two-systems account of Daniel Kahneman (2011) has it, most everyday behaviour runs on the fast, automatic, intuitive "System 1" rather than the slow, effortful, reasoning "System 2". A nudge works on System 1: it sets small, cheap cues in the environment that trigger behaviour below the level of conscious deliberation. Its goal is statistical, to shift the average outcome, not to determine any single decision. Thaler and Sunstein defend this as **libertarian paternalism**: paternalist because it steers people toward what is held to be good for them, libertarian because freedom of choice is formally preserved. The tension in that phrase is the whole ethical debate.

DEFINITION

Nudge and choice architecture

A nudge is a feature of the *choice architecture* that predictably steers behaviour without forbidding options or changing incentives; the classic lever is the *default*.

6.2 From nudges to digital nudges and dark patterns

In interface design the nudge is everywhere. Consent dialogues make the "accept all" button large and colourful and the "reject" option small and grey; travel insurance is pre-ticked when you book a flight; the whole user journey is arranged to maximise engagement and conversion. Here the ethical stakes sharpen, because the declared goal is often the *provider's* interest, not the user's, and the line between a benevolent nudge and a manipulative **dark pattern** turns on exactly that. A nudge that helps a user decide as they themselves would on reflection respects autonomy; one engineered to extract a click or a subscription against the user's considered interest exploits System 1 to override it. This is more than an ethical worry: it is why the EU AI Act's very first

prohibition targets AI that uses subliminal or manipulative techniques to distort behaviour and cause harm, and it connects directly to the regulatory units later in the course.

QUICK CHECK

A cookie banner shows a big highlighted “Accept all” button and hides “Reject” behind an extra menu. What is this, in the vocabulary of this unit?

A neutral presentation of equal options
A digital nudge that risks becoming a manipulative dark pattern
A form of machine agency
A moral patient

Because choice architecture is inescapable, the ethically decisive question is not whether a design steers but how far it steers and against whom. Design the consent screen of an AI assistant yourself in the explorer below and watch when a legitimate nudge tips over into a manipulative dark pattern.

DEEP DIVE Deep dive: is libertarian paternalism really libertarian?

Critics argue the phrase papers over a genuine loss of autonomy. Because choice architecture is inescapable, the honest question is not *whether* to steer but *toward what* and *by whom*, and a default set by a self-interested platform is paternalism without the benevolence. Defenders reply that since some arrangement is unavoidable, choosing the one that most people would endorse on reflection is the least bad option, and that transparency plus an easy opt-out preserves real freedom. The disagreement maps neatly onto the schools of the previous unit: a consequentialist weighs the improved average outcome, a Kantian asks whether nudging through System 1 treats the person as a rational end or as a mechanism to be triggered. For AI systems, whose scale and personalisation make nudges vastly more powerful than a salad bar, this is not an academic quarrel.

07
KAPITEL

From agency to responsibility

This unit has pulled the human-AI relation apart into its components. We separated moral status, patiency and agency; met the machine-ethics claim that AI can be accountable without being responsible; saw in Heider and Simmel and the intentional stance why we over-attribute mind to machines; followed Actor-Network Theory in spreading agency across humans and things; recognised AI as a sociotechnical system that harvests human computation and shapes us through affordances; and watched choice architecture steer behaviour through nudges. Running

through all of it is one unresolved thread: AI can be the source of morally weighty action, yet it cannot answer for it. That gap between causation and accountability is the subject of the next unit, which asks who is responsible when AI causes harm, and confronts the responsibility gap head-on.

08
KAPITEL
References

8.1 Literature

- Coeckelbergh, M. (2020): *AI Ethics*. MIT Press, Cambridge, MA.
- Floridi, L. & Sanders, J. W. (2004): On the Morality of Artificial Agents. *Minds and Machines* 14, pp. 349-379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d
- Heider, F. & Simmel, M. (1944): An Experimental Study of Apparent Behavior. *The American Journal of Psychology* 57(2), pp. 243-259. https://doi.org/10.2307/1416950
- Kahneman, D. (2011): *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- Latour, B. (1992): Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts, in: Bijker, W. E. & Law, J. (eds.), *Shaping Technology / Building Society*. MIT Press, Cambridge, MA, pp. 225-258.
- Matthias, A. (2004): The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology* 6(3), pp. 175-183. https://doi.org/10.1007/s10676-004-3422-1
- Thaler, R. H. & Sunstein, C. R. (2008): *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, New Haven.
- von Ahn, L. & Dabbish, L. (2004): Labeling Images with a Computer Game, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. ACM, New York, pp. 319-326. https://doi.org/10.1145/985692.985733

8.2 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), Art. 5(1)(a). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689