thws Technical University of Applied Sciences Würzburg-Schweinfurt

FAKULTÄT INFORMATIK UND WIRTSCHAFTSINFORMATIK

LECTURE NOTES

Ethics 101: Foundations of Moral Philosophy

Values and norms, the is/ought gap, and the three schools of ethics

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 clarified what "AI" means (at least to us) and why its social effects call for ethical scrutiny at all; this unit supplies the conceptual toolkit that the rest of the course will rely on. The aim is not to turn you into moral philosophers, but to give you an instrument with which you can recognise, justify and articulate ethical requirements for AI systems. When a recommender system shapes what millions read, when a scoring model decides who gets a loan, or when a chatbot advises a user in distress, the question is no longer only whether the system *works*, but whether its use is *right*. That second question cannot be answered with engineering alone, and answering it well begins with a small number of distinctions that moral philosophy has refined over two and a half thousand years.

01
KAPITEL

What is ethics? Basic concepts

1.1 Ethics, morality and right action

The word *ethics* derives from the ancient Greek *ethos* (custom, character), yet it is not identical with it. Ethics is the **study of right action**, the systematic, scholarly inquiry into what is good and right and into how we ought to act. Where *morality* denotes the stock of value convictions a group actually lives by, ethics as a philosophical discipline asks after the **justification** of those convictions. It does not hand out ready-made answers; it supplies procedures by which answers can be defended. In an academic context, ethics is therefore synonymous with **moral philosophy**: the systematic reflection on the values and norms of individual and collective action.

For AI this shift of perspective is decisive. Technical systems increasingly take or support decisions that affect people, and the question whether a model is *accurate* can be settled empirically, whereas the question whether its deployment is *justified* cannot. A face-recognition system may reach 99% accuracy and still be wrong to deploy. Recognising that these are two different questions, answered by two different methods, is the first thing ethics teaches an engineer.

DEFINITION

Ethics vs. morality

Morality is the lived stock of value convictions of a group; ethics is the scholarly reflection on their justification and validity.

1.2 Values, principles, norms: a ladder of concepts

The basic concepts of ethics build on one another like rungs of a ladder. **Values** are action-guiding standards of orientation. On the prevailing view in philosophy they are not properties of facts or objects but *claims to validity* that human beings raise; the question is therefore not whether a value exists but whether it is valid and for whom. The ethical concept of value must also be kept

apart from the economic notion of exchange value. A decision is not ethical merely because it is profitable, a distinction that matters acutely in a field where the most lucrative data practice is rarely the most respectful one.

Principles join a value to a first "ought" statement: value X *shall* obtain. Human dignity, for instance, grounds the principle that every person be treated with respect. **Norms**, finally, are more or less binding "ought" sentences for a group or an individual, ranging from custom and convention through morality to law, and graded by their degree of bindingness and social control. A norm is *specific to its addressees*; a principle is typically **universal** (it applies to all persons, or all persons in a certain role) and **timeless** (it is not limited to a specific historical period). When a principle is made concrete for its addressees, we speak of a **rule**. Norms *can* be traced back to principles and, through them, to values, but they need not be: a paper-size standard such as DIN A4 is a norm without any recognisable ethical value behind it.

EXERCISE · TERMS

A **value** is an action-guiding standard of orientation, a **principle** links it to a first "ought" statement, and a concrete, addressee-specific behavioural norm is called a **rule**.

02
KAPITEL

The is/ought gap and the naturalistic fallacy

Before we survey the schools of ethics, one logical guard-rail has to be in place, because almost every sloppy argument about AI trips over it. In the *Treatise of Human Nature* (1739/40), David Hume observed that authors slide imperceptibly from statements about what *is* to statements about what *ought* to be, without ever explaining how the second follows from the first. This is the **is/ought gap**: no set of purely descriptive premises can, by logic alone, entail a normative conclusion. To infer an "ought" directly from an "is" is what later philosophy came to call the **naturalistic fallacy**.

The point is not academic hair-splitting. Consider the invalid inference: *fact*, people are starving in country X; *conclusion*, therefore they must be helped. The conclusion may be true, but it does not follow, because the premises contain no value. The valid version makes the normative premise explicit: *premise*, human dignity includes freedom from hunger; *fact*, people in country X are starving; *conclusion*, therefore they must be helped. The fallacy is avoided by always stating the value or rule that carries the "ought" openly, rather than smuggling it in under cover of a fact.

	Fallacious inference	Correct inference
Normative premise	(missing)	Human dignity includes freedom from hunger.
Descriptive fact	People are starving in country X.	People are starving in country X.
Conclusion	Therefore they must be helped.	Therefore they must be helped.

For AI ethics this discipline is indispensable, because the field is full of hidden "ought"s dressed as facts. "The model is more accurate than human judges, so we should use it" is a naturalistic fallacy: accuracy is a fact, "should use" an ought, and the missing premise (that we ought to maximise accuracy even at the cost of, say, contestability) is exactly the ethical claim that deserves scrutiny. **Naming the normative premise is where honest AI ethics starts.**

Train your eye on the distinction in the detector below: for each statement, decide whether it is descriptive, normative, or a fallacious leap from is to ought, and check your reasoning against Hume's law.

QUICK CHECK

"Everyone already shares their data online, so it is fine for us to scrape it for training." Which error does this argument commit?

A valid utilitarian calculation
The naturalistic fallacy: it derives an "ought" from a mere "is" without stating the value premise
A category mistake about moral status
A breach of the categorical imperative

03
KAPITEL

The three schools of ethics

Western ethics knows three great schools of thought, distinguished by *what* they use to measure right action: the intention behind it, its consequences, or the procedure by which its guiding rules are set. They are not mutually exclusive camps but complementary lenses, and mature practical judgement usually draws on all three.

Virtue ethics (an *ethics of conviction*) asks whether the *intention* and character behind an action match certain values; its goal is the good, flourishing life. Its founding figure is *Aristotle* (384-322 BC), who in the *Nicomachean Ethics* distinguishes intellectual from moral virtues and locates each moral virtue as a **mean** between two vices (courage, for instance, between cowardice and recklessness). The Jewish and Christian traditions later added the theological virtues of faith,

hope and love. Applied to AI, virtue ethics asks less "which rule was broken?" than "what kind of practice, and what kind of engineer or organisation, does this system cultivate?"

Utilitarianism (an *ethics of consequences* or *responsibility*) measures actions by their *outcomes*: right is what maximises utility for the greatest number. On the classical formulations of Jeremy *Bentham* (1748-1832) and John Stuart *Mill* (1806-1873), utility spans material advantage, pleasure and the avoidance of suffering, and the interests of all count equally, one's own included. Utilitarianism is therefore expressly **not** an egoistic doctrine. It supplies the natural language of cost-benefit reasoning about AI, from "data donations" that improve a model for everyone to the triage calculus of an emergency system.

Deontology, the ethics of duty founded by Immanuel *Kant* (1724-1804), grounds the "ought" in practical reason alone. By virtue of their will, human beings can impose duties on themselves (**autonomy**), and the substantive test is the **categorical imperative**, in its universalisation formula ("act only according to that maxim whereby you can at the same time will that it should become a universal law") and its humanity formula (never treat a human being *merely* as a means, but always also as an end). Because it is procedural rather than outcome-driven, deontology is the natural home of AI ethics *codes* and *charters*, which try to fix universalisable duties for all developers in advance.

Explore in the widget below how the three schools judge the same AI dilemma differently.

QUICK CHECK

A safety engineer decides to disclose a model flaw because doing so "produces, on balance, the greatest benefit for everyone affected". Which school does this reasoning follow?

Virtue ethics
Utilitarianism
Deontology
Machine ethics

The three schools are not sealed compartments. In practice they yield three complementary rules of thumb: act *virtuously* in the sense of the doctrine of virtue, act so as to produce the *greatest benefit for the most people*, and act in accordance with the *duties* that a fair procedure has established. Where they collide, the collision itself is usually the ethical heart of the case, as the following study shows.

CASE STUDY

Case: the utilitarian triage model

A hospital pilots an AI that allocates scarce ICU beds. In a mass-casualty scenario it would deny a bed to one patient with poor prognosis in order to save five others, maximising expected lives saved. The numbers are sound. Should the hospital deploy it as an autonomous decision rule?

Solution. The model embodies a clean **utilitarian** calculus: five lives outweigh one, so the aggregate outcome is better. Yet **deontology** sets a hard limit exactly here. Deciding, by rule,

to sacrifice one identified patient as a mere instrument for others treats that person *merely as a means*, in breach of the humanity formula of the categorical imperative. **Virtue ethics** adds a third worry: a clinician who defers to a cold optimiser may erode the very disposition of care that medicine depends on. The lesson is not that consequences are irrelevant but that a defensible design keeps the human decision in the loop, uses the model to *inform* rather than *replace* judgement, and documents openly which value premise (aggregate benefit vs. the inviolability of the individual) is doing the work. This is also why later units treat "human oversight" as a legal, not merely ethical, requirement.

Deontology tells us to act on universalisable maxims, but who decides *which* maxims are valid? A later unit on the public sphere and disinformation takes up exactly this question through **discourse ethics**, the German tradition developed by Jürgen *Habermas* and Karl-Otto *Apel*, which answers it procedurally rather than from a single armchair.

04

KAPITEL

How moral judgement develops: Kohlberg

The schools above concern the *justification* of norms. A different, empirical question is how the human capacity for moral judgement actually *develops*. The psychologist Lawrence **Kohlberg** studied this by presenting people with dilemmas (most famously the "Heinz dilemma", whether a man may steal an over-priced drug to save his dying wife) and analysing not the answer but the *reasoning* behind it. He found that moral reasoning matures through six stages grouped into three levels, from an orientation toward punishment and reward, through conformity to social rules and law, up to reasoning from self-chosen universal principles.

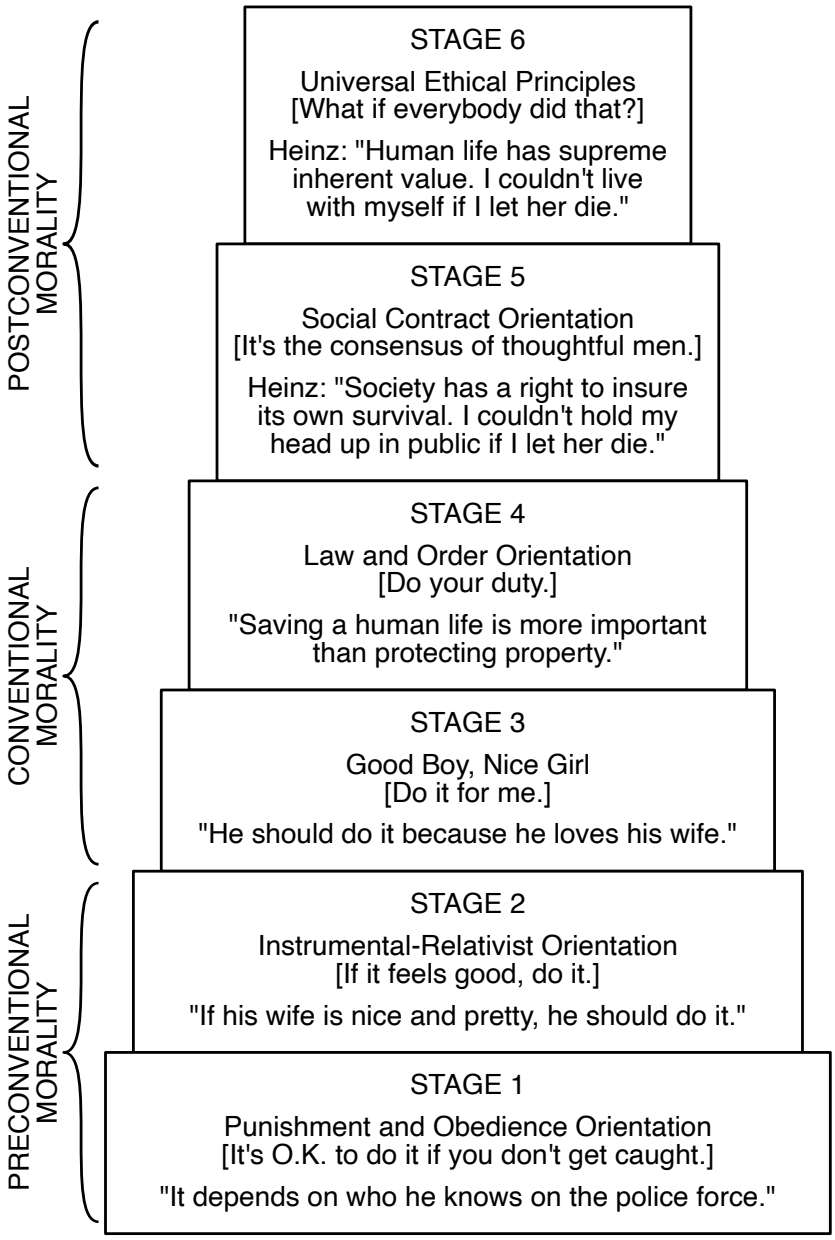

Figure 1: Kohlberg's stages of moral development: reasoning matures from the preconventional level (obedience and self-interest), through the conventional level (approval and law-and-order), to the postconventional level (social contract and universal ethical principles), illustrated with responses to the Heinz dilemma (source: Kohlberg Model of Moral Development, Lawrence Kohlberg / Em Griffin / cmglee, Wikimedia Commons, CC BY-SA 4.0).

Two features of the model matter for AI ethics. First, higher stages reason from *principles* rather than from what is rewarded or merely customary, which mirrors the move from "the model is legal" or "everyone does it" to "is this defensible on universal grounds?". Second, and more provocatively, an AI system trained on human data absorbs the *distribution* of moral reasoning in that data, which sits mostly at the conventional level, not at the principled top. A model that simply reproduces prevailing conventions has no route to the postconventional critique that ethics ultimately demands, which is one reason automated moral judgement cannot be left to imitation alone.

QUICK CHECK

A user justifies a design choice by saying “it is standard practice in the industry and nobody complains”. Which of Kohlberg’s levels does this reasoning exemplify?

Postconventional (universal principles)
Conventional (conformity to social rules and expectations)
Preconventional (avoiding punishment)
It lies outside Kohlberg’s model

05
KAPITEL

Two systems of thinking: why ethics needs the slow one

A final building block comes from cognitive psychology. In *Thinking, Fast and Slow* (2011), Daniel **Kahneman** popularised the distinction between two modes of thought. **System 1** is fast, automatic, intuitive and emotional: it governs most day-to-day behaviour and produces snap judgements with little effort. **System 2** is slow, effortful, deliberate and rule-following: it is the mode of explicit reasoning and conscious choice. The two are not rival theories of the mind but a useful map of when we coast on intuition and when we actually think.

Ethical reasoning is quintessentially System 2 work. Spotting a naturalistic fallacy, weighing a utilitarian calculus against a deontological limit, or checking whether all affected parties were heard requires the slow, effortful system, precisely the system that a busy engineering sprint tends to skip. This has a double relevance for AI. On the developer’s side, ethical review must be built into the process so that System 2 is engaged deliberately rather than left to a hurried System 1. On the user’s side, AI interfaces increasingly *exploit* System 1 through defaults, framing and nudges, a theme the next unit takes up directly. Knowing which system is in play is thus both a tool of self-discipline for designers and a lens for critiquing the systems they build.

DEEP DIVE Deep dive: free will, evolutionary ethics and the limits of “is”

If our choices are, as some neuroscientists argue, largely determined by genes, neurons and situation, can we be held responsible at all? “Evolutionary ethics” treats moral behaviour as a pre-programmed disposition shaped by natural selection. Whatever its descriptive merits, note that it cannot, on its own, settle a normative question without committing the naturalistic fallacy: that a disposition *evolved* does not entail that acting on it is *right*. The debate over free will matters here because responsibility (a later unit) presupposes some control and awareness on the agent’s part. It also frames a hard question about AI: a system

with neither will nor consciousness cannot be a moral agent in the full sense, however fluently it produces moral-sounding text.

06
KAPITEL

From concepts to constellations

With this toolkit in hand, values and norms, the is/ought gap, the three schools, and the psychology of moral judgement, we can now ask the questions that make AI ethics distinctive. Chief among them is whether, and in what sense, an artificial system can be said to *act* at all, and how humans and machines share the moral stage. That is the subject of the next unit on agency and the human-AI relation.

07
KAPITEL

References

7.1 Literature

- Aristotle (2009): *The Nicomachean Ethics* (translated by David Ross, revised by Lesley Brown). Oxford University Press, Oxford.
- Bentham, J. (1789/1996): *An Introduction to the Principles of Morals and Legislation* (edited by J. H. Burns & H. L. A. Hart). Clarendon Press, Oxford.
- 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
- Hume, D. (1739-40/2000): *A Treatise of Human Nature* (edited by David Fate Norton & Mary J. Norton). Oxford University Press, Oxford.
- Kahneman, D. (2011): *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- Kant, I. (1785/1997): *Groundwork of the Metaphysics of Morals* (translated and edited by Mary Gregor). Cambridge University Press, Cambridge.
- Kohlberg, L. (1981): *The Philosophy of Moral Development: Moral Stages and the Idea of Justice*. Harper & Row, San Francisco.
- Mill, J. S. (1863/1998): *Utilitarianism* (edited by Roger Crisp). Oxford University Press, Oxford.