thws Technical University of Applied Sciences Würzburg-Schweinfurt

FAKULTÄT INFORMATIK UND WIRTSCHAFTSINFORMATIK

LECTURE NOTES

AI Ethics Guidelines

Convergent principles, blind spots and the legal quality of soft law

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 watched AI reshape work and strain the planet's resources, and we asked who should steer a transformation of that scale. This unit turns to the first collective answer the world reached for: a wave of ethics guidelines issued by international organisations. Between roughly 2016 and 2021 governments, expert bodies and industry produced document after document promising "trustworthy", "responsible" or "human-centric" AI. The task of this unit is to read that wave critically. We will see where the guidelines converge, what they systematically leave out, why critics speak of "ethics washing", and, crucially, what these instruments are worth in legal terms. Along the way we build a working matrix of ethical principles that will serve us for this course.

01
KAPITEL

The wave of guidelines and its convergent canon

1.1 From a flood of documents to a systematic reading

The years after 2016 saw an extraordinary proliferation of AI ethics guidelines. Tech firms, national governments, civil-society organisations and international institutions all published their own charters in quick succession, so many that a comparative science of the field became necessary simply to see the wood for the trees. A early but sound systematic scholarly evaluation came from Thilo Hagendorff, whose 2020 study *The Ethics of AI Ethics* analysed twenty-two guidelines drawn from across this landscape, from Google, Microsoft and IBM through the OECD, the European Commission and the G7 to the Future of Life Institute and the IEEE. His method was straightforward and revealing: which ethical themes appear, how often, with what degree of commitment, and, above all, whose interests do the chosen principles protect?

The first finding is convergence. Beneath the differences of wording, the guidelines cluster around a remarkably stable canon of principles. Transparency and explainability appear almost universally, closely followed by fairness and non-discrimination, privacy and data protection, safety and technical robustness, and accountability. Human oversight and an appeal to the common good round out the set. This overlap is genuinely useful: it gives us a shared vocabulary and suggests a rough international consensus about what "good" AI should respect. But convergence alone proves little, because a principle that everyone can sign is often a principle that binds no one.

DEFINITION

The convergent canon (Hagendorff 2020)

Across twenty-two guidelines a stable set of principles recurs: transparency and explainability, fairness and non-discrimination, privacy, safety and robustness, accountability, human oversight, and orientation towards the common good.

1.2 What the guidelines leave out

Far more instructive than the shared canon, for Hagendorff, is the pattern of *absence*. Certain questions are missing with a regularity that is itself an argument. The employment effects of AI, automation, deskilling and working conditions, are barely addressed, although they touch the interests of workers rather than firms. Questions of power, who controls AI systems, who profits and how concentration among a few companies might be checked, are almost entirely absent, which is unsurprising given that many guidelines were written by exactly those companies. The environmental cost of training large models, the militarisation of AI and the dynamics that arise when autonomous systems interact with one another are likewise underplayed or ignored. Read against its silences, the consensus looks less like a neutral distillation of ethics and more like the set of commitments that were comfortable to make.

EXERCISE · TERMS

The guidelines converge on a shared **canon** of principles such as transparency and fairness, yet systematically omit questions of **power**, of labour and of the environment, which is why their consensus can be misleading.

QUICK CHECK

According to Hagendorff (2020), which theme is systematically *underrepresented* across the guidelines he analysed?

Transparency and explainability
The concentration of power among a few AI companies
Fairness and non-discrimination
Privacy and data protection

02
KAPITEL

Ethics washing and the enforcement gap

The convergence of principles and the pattern of omissions point to a deeper structural problem, and it is here that Hagendorff makes his sharpest intervention. Nearly all the guidelines he studied share three features: they are voluntary, they are non-binding, and they come without any external mechanism of enforcement. There are typically no independent audits, no sanctions for breach, no external validation of how the principles are drafted, and no seat at the table for the groups most affected, users, workers and vulnerable populations. A firm can subscribe to a charter of admirable principles and change nothing about its conduct.

Hagendorff names this pattern **ethics washing**, in deliberate analogy to greenwashing. Ethics guidelines perform a legitimation function towards the public, regulators and investors without producing any substantive change in behaviour or any binding commitment. The point is not that every guideline is cynical, but that the *form* itself, self-imposed, unaudited and unenforceable, invites the suspicion.

The critique has since become mainstream, and it carries a clear regulatory implication: ethics guidelines can *complement* law but cannot *replace* it. Without binding rules, principles remain rhetoric. This is precisely the gap that the EU AI Act was later designed to close, hardening principles such as transparency, human oversight and robustness into enforceable obligations for high-risk systems, a bridge we will cross in a later unit.

DEFINITION

Ethics washing

The use of voluntary, unaudited and unenforceable ethics guidelines as a public-relations and legitimation device, projecting responsibility without producing binding self-commitment or behavioural change.

CASE STUDY

Case: a corporate “AI ethics charter”

A large technology company publishes a glossy “Responsible AI Charter” committing itself to fairness, transparency and human dignity. It appoints an internal ethics board whose members it selects and pays, publishes no audit results, and provides no route for affected users to lodge complaints. Critics call this “ethics washing”. Are they right, and what would distinguish a genuine commitment?
Solution. The critics have a strong case on Hagendorff’s own criteria. The charter is voluntary, self-drafted and unaudited; the ethics board lacks independence because the company

appoints and pays it; there is no sanction for breach and no participation by those affected. The charter therefore serves a legitimation function towards the public and regulators without binding the firm to anything. What would distinguish a genuine commitment is external accountability: independent audits with published results, effective grievance mechanisms for affected persons, participation of workers and user groups in drafting, and consequences for non-compliance. Failing that, the alternative is not better self-regulation but binding hard law, which is exactly the direction European regulation subsequently took.

03

KAPITEL

The legal quality of soft law

To evaluate the guidelines fairly we need a sharper instrument than the everyday word “non-binding”. Lawyers distinguish **hard law** from **soft law**. Hard law is enforceable with the power of the state: a regulation, a directive transposed into national law, a court judgment. Soft law comprises instruments that articulate standards and expectations but cannot, by themselves, be enforced against those who ignore them: recommendations, declarations, codes of conduct, expert guidelines. The distinction is not the same as “important” versus “unimportant”. Soft law can shape behaviour powerfully through peer pressure, reputation and diffusion, and it often serves as the drafting bed for later hard law. But when a firm decides to disregard it, no authority can compel compliance and no fine follows.

Two further gradations matter for the frameworks in this unit. First, an instrument’s *addressee* varies: some guidelines speak to states (recommendations to be reflected in national policy), others directly to organisations that build AI (codes of conduct). Second, and decisively, a single instrument type stands apart from all the others: the international **treaty** or **convention**. A convention is soft only until a state signs and ratifies it; upon ratification it becomes binding international law for that state, which must then give it domestic effect. This is why, among the frameworks profiled below, the Council of Europe’s instrument occupies a category of its own. Use the comparator to place each framework on the spectrum from soft law to binding treaty before we profile them one by one.

DEFINITION

Hard law vs. soft law

Hard law (regulations, directives, judgments) is enforceable with state power; soft law (recommendations, declarations, codes) sets standards but cannot, on its own, be enforced. A treaty is soft only until a state ratifies it, whereupon it becomes binding.

04
KAPITEL

Five frameworks in profile

4.1 European Union: Ethics Guidelines for Trustworthy AI (2019)

The European Union is a supranational organisation with the power to legislate directly for its member states, but the 2019 guidelines are deliberately not such legislation. They were produced by a **High-Level Expert Group on AI**, convened by the Commission and drawn from computer science, philosophy, law, industry and civil society, and finalised in 2019 through no formal adoption procedure. Their legal quality is that of non-binding guidelines. Their importance lies elsewhere: they crystallised the European vocabulary of "Trustworthy AI" and became the principal reference point for the later AI Act.

Substantively the guidelines rest on four ethical principles: respect for human autonomy, prevention of harm, fairness, and explicability. From these they derive seven concrete requirements, among them human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal and environmental well-being, and accountability. Because this framework supplies the human-oversight and traceability concepts we examine below, and because it seeds the AI Act, it repays closer attention than the others.

4.2 OECD: Recommendation on Artificial Intelligence (2019)

The Organisation for Economic Co-operation and Development groups thirty-eight mostly industrialised economies and works chiefly through studies and recommendations. Its **Recommendation on Artificial Intelligence** (OECD/LEGAL/0449) was adopted in May 2019 and updated in 2023 and 2024. Its legal quality is that of an intergovernmental standard, a Council Recommendation that members are expected to reflect in national policy but that binds no one. It sets out five value-based principles: inclusive growth and well-being, human-centred values and fairness, transparency and explainability, robustness, security and safety, and accountability. Its most consequential legacy is definitional rather than principled: the OECD's definition of an AI system was adopted almost verbatim by the EU AI Act, which lends this piece of soft law an unusual measure of indirect legal force.

4.3 UNESCO: Recommendation on the Ethics of AI (2021)

UNESCO, the UN's specialised agency for education, science and culture with 194 member states, adopted its **Recommendation on the Ethics of Artificial Intelligence** in November 2021. It is the first genuinely global standard-setting instrument on AI ethics. As a Recommendation addressed to member states it is not a treaty and not legally binding, but it stands out for its breadth and for one procedural innovation. Its principles include proportionality and doing no harm, human oversight and determination, fairness and non-discrimination, sustainability, and the right to privacy. The innovation is accountability of a soft kind: member states report on implementation every four years, and the text is operationalised through readiness and impact

assessment methodologies. That reporting cadence gives the Recommendation more traction than a one-off declaration, even though it carries no sanction.

4.4 G7: Hiroshima Process International Code of Conduct (2023)

The G7 is an informal coalition of seven industrialised states, with the EU as observer. Its **Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems** emerged from the 2023 Hiroshima Summit. Two features set it apart. Its legal quality is that of voluntary, non-binding guidance, with no monitoring body and no sanctions. And its addressee is unusual: rather than speaking to states, it speaks directly to the organisations building frontier AI, asking them to identify and mitigate lifecycle risks, report publicly and transparently, share incident information responsibly, and invest in security and content authentication. It is thus a bridge to the self-regulation we examine next: a code that firms are invited to adopt, but only if they choose.

4.5 Council of Europe: Framework Convention on AI (2024)

The Council of Europe must not be confused with any EU body. It is an international organisation of forty-six member states, seated in Strasbourg, best known for the European Convention on Human Rights and its Court. Alongside recommendations, it drafts binding international treaties, and its **Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law**, opened for signature in 2024, is precisely that. This is the decisive difference from every other instrument in this unit. As a convention it is soft only until signed and ratified; once a state ratifies it, it becomes **binding international law** for that state, which must give its obligations domestic effect. Its principles centre on the protection of human rights, the integrity of democratic processes and the rule of law, transparency and oversight, accountability, and equality. Its real-world reach nonetheless depends on how many states ratify it and how faithfully they implement it, so even a binding treaty is no automatic guarantee of effect.

DEEP DIVE Deep dive: why the Council of Europe convention is the outlier

Every other framework here is soft law: guidelines, recommendations or a voluntary code, none enforceable against a body that ignores them. The Framework Convention is a treaty, and treaties change legal quality on ratification. That distinction is not merely formal. It means that, for a ratifying state, the convention's guarantees can in principle be invoked and must be reflected in domestic law, whereas the OECD or UNESCO texts remain, however influential, standards without teeth. The catch is participation: a convention binds only its parties, so a treaty ratified by few states may in practice reach fewer actors than a widely diffused recommendation. Legal form and factual reach do not always run together.

05
KAPITEL

Human oversight and traceability in practice

Two of the convergent principles deserve concrete treatment, because the EU guidelines translate them into operational forms that recur throughout AI regulation. Both answer questions we raised in the earlier units on responsibility and explainability, and both reappear as hard-law duties in the AI Act.

5.1 Modes of human oversight

The principle of respect for human autonomy demands that AI systems be designed with human oversight in mind, so that people can intervene to prevent harm and preserve accountability. The guidelines distinguish three modes by how tightly the human is coupled to the decision.

In **human-in-the-loop (HITL)** the human is an active participant: every decision the system proposes is reviewed and approved by a person before it takes effect.

In **human-on-the-loop (HOTL)** the human is not part of each decision but supervises the system in operation and can intervene when something goes wrong.

In **human-in-command (HIC)** the human retains overall authority over the system, deciding whether, when and in what contexts it is deployed at all, and able to override or shut it down.

The three modes trade immediacy of control against scalability, and the right choice depends on the stakes and the tempo of the decision. Where actions unfold on inhuman timescales, an oversight arrangement that cannot actually intervene in time is oversight in name only, which is why the choice of mode is itself an ethical decision.

DEFINITION

HITL, HOTL, HIC

Human-in-the-loop: a person approves each decision before execution. Human-on-the-loop: a person supervises operation and can intervene. Human-in-command: a person retains overall authority to deploy, override or shut the system down.

EXERCISE · TERMS

When a person must approve every single output before it takes effect, the system runs **human-in-the-loop**; when a person merely monitors an operating system and steps in on problems, it runs **human-on-the-loop**; when a person decides whether the system is deployed at all, they exercise **human-in-command**.

QUICK CHECK

A bank's AI proposes loan decisions, and a caseworker must review and sign off on each one before it is communicated to the customer. Which mode of oversight is this?

Human-in-the-loop (HITL)
Human-on-the-loop (HOTL)
Human-in-command (HIC)
No human oversight

Choosing the right mode is not a matter of taste but of stakes and tempo, and every mode brings its own characteristic way of failing. Work through the use cases below to see which oversight mode fits each situation, what the human is actually expected to do, and the failure risks the mode invites, from rubber-stamping and automation bias to intervening too late to matter.

5.2 Traceability, recognisability and explainability

The principle of transparency likewise breaks down into practical requirements. **Traceability** asks that the datasets and the processes yielding a system's decisions, including data gathering, labelling and the algorithms used, be documented to the best possible standard, so that outcomes can be reconstructed after the fact.

Recognisability holds that a system should not pass itself off as human: people have a right to know when they are interacting with an AI.

Explainability requires that the reasons for a decision can be communicated to those affected, whether through technical methods of explainable AI or through plain-language account.

These three are not the same. Traceability serves the internal ability to audit and reconstruct; explainability serves the external duty to answer to the people a decision touches. Together they operationalise the answerability we grounded in the unit on responsibility, and they resurface, as we will see, as concrete documentation, transparency and oversight duties in the AI Act.

DEEP DIVE Deep dive: from a soft principle to a hard duty

The journey of "human oversight" illustrates the whole arc of this unit. It begins as an ethical aspiration in the 2019 guidelines, one soft-law principle among many, unenforceable on its own. It is echoed across the OECD, UNESCO and other texts, gaining the weight of international consensus but still no teeth. It is finally hardened into a binding legal obligation for high-risk systems in the EU AI Act, complete with the possibility of severe fines for non-

compliance. The same path, from aspiration through consensus to enforceable rule, is exactly what Hagendorff's critique of ethics washing recommends: principles matter, but only binding law closes the enforcement gap.

06
KAPITEL

Towards a working matrix, and on to self-regulation

We can now assemble the payoff of this unit. Reading the guidelines together yields a workable matrix of ethical principles, transparency, fairness, privacy, safety, accountability, human oversight and the common good, that we can carry forward as a checklist for assessing any AI system, while remembering the systematic blind spots, power, labour and environment, that the consensus tends to hide. We have also learned to read these instruments in legal terms, distinguishing enforceable hard law from soft law that persuades but cannot compel, with the Council of Europe's convention as the one framework that crosses into binding international law once ratified. Hagendorff's warning frames the whole picture: without enforcement, ethics risks becoming public relations. That warning sets up the next unit directly, for it asks what happens when firms move from signing international guidelines to writing their own internal codes of ethics, and whether such self-regulation can ever be more than ethics washing by another name.

07
KAPITEL

References

7.1 Literature

- Hagendorff, T. (2020): *The Ethics of AI Ethics: An Evaluation of Guidelines*. Minds and Machines 30(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8
- Jobin, A., Ienca, M. & Vayena, E. (2019): *The Global Landscape of AI Ethics Guidelines*. Nature Machine Intelligence 1, 389-399. https://doi.org/10.1038/s42256-019-0088-2

7.2 Norms & Standards

- High-Level Expert Group on Artificial Intelligence (2019): *Ethics Guidelines for Trustworthy AI*. European Commission, Brussels. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
- OECD (2019, amended 2023 and 2024): *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449. OECD, Paris. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
- UNESCO (2021): *Recommendation on the Ethics of Artificial Intelligence*, SHS/BIO/REC-AIETHICS/2021. UNESCO, Paris. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics
- G7 (2023): *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems*. https://digital-strategy.ec.europa.eu/en/library/hiroshima-process-international-code-conduct-advanced-ai-systems
- Council of Europe (2024): *Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law*, CETS No. 225. Council of Europe, Strasbourg. https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence