thws Technical University of Applied Sciences Würzburg-Schweinfurt

FAKULTÄT INFORMATIK UND WIRTSCHAFTSINFORMATIK

LECTURE NOTES

Internal Codes & Self-Regulation

Regulatory choice, corporate codes of ethics and the standards that operationalise them

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

The previous unit closed with Hagendorff's warning that ethics guidelines issued from the outside can slide into public relations, so this unit turns inward and asks whether firms writing their own rules can do any better. Before we can judge internal codes of ethics, however, we need a map of the regulatory landscape as a whole. Why does a liberal state so often prefer that companies regulate themselves? What exactly does a corporate code of ethics achieve, and for whom? And what stands between a lofty principle and an engineer's actual workflow? This unit answers those questions in three moves: it lays out the forms of regulation and the logic of regulatory choice, it dissects internal codes and OpenAI's Preparedness Framework as a live specimen, and it introduces the technical standards that try to turn ethics into repeatable process.

01
KAPITEL

The forms of regulation

1.1 Two poles: state regulation and self-regulation

At its simplest, the field of regulation stretches between two poles. **State regulation** means that the state binds private actors in order to reach public-interest goals. Its instruments are laws and the directives of authorities, and its distinctive feature is that it is enforceable by the coercive power of the state: a firm that ignores a binding rule can often be fined, services can be prohibited or products withdrawn from the market. **Self-regulation** means that private actors regulate themselves to reach those same goals. Here the state holds back, trusting that societal or market processes will deliver the desired outcome, and the instruments are codes of conduct, professional standards and guidelines, backed only by private enforcement mechanisms such as expulsion from an association or the loss of reputation.

The difference is not that one is serious and the other decorative. Self-regulation can steer behaviour powerfully, and it often moves faster and with more technical insight than any ministry could. But it rests on a fragile premise: that the self-interest of the regulated actors runs with the public goal rather than against it. Where a firm profits from exactly the conduct the rule is meant to curb, self-regulation predictably fails, because no private mechanism compels a company to act against its own bottom line.

DEFINITION

State regulation vs. self-regulation

State regulation binds private actors through law, enforceable by state coercion; self-regulation has private actors bind themselves through codes and standards, enforceable only by private means such as reputation or membership sanctions.

1.2 Forms of state regulation, graded by intensity

State regulation is itself not one thing. Anthony Ogus (2004) sorts its instruments along a scale of **regulatory intensity**, from the lightest touch on the addressee's freedom to the heaviest. At the gentle end sits **transparency regulation**, which only requires actors to disclose information and otherwise leaves their conduct untouched. Then come **private regulation**, which works by shaping private-law relations and liability, and **incentive regulation**, which steers behaviour through taxes, levies or subsidies rather than direct commands. Heavier still is **command-and-control regulation**, the classic model of binding prohibitions and requirements policed by an authority. At the most intrusive end lies **prior approval**, where an activity may not begin at all until the state has licensed it. The ordering matters because it gives the regulator a menu ranked by how deeply each option cuts into the freedom of those it governs.

EXERCISE · TERMS

A rule that merely forces a firm to **disclose** how its system works is a light-touch **transparency** regulation, whereas a rule that forbids market entry until the state grants a licence is the far more intrusive **prior approval**.

1.3 Why liberal states lean towards the lightest tool

The grading is not academic. The constitutional order of liberal jurisdictions, with binding fundamental rights, effectively forces a regulator to choose the least intrusive form that is still capable of reaching the public goal. This is the logic of proportionality: if the chosen instrument is more intrusive than necessary, it violates the fundamental rights of the addressees, and the law or directive can be struck down in court. Alongside the constitutional constraint runs a cultural one. The political common sense of liberal democracies with market economies pushes in the same direction, keeping state intervention in society, and above all in markets, to the minimum required. These two pressures together explain the structural preference for self-regulation and for the lightest workable tool: not because self-regulation is always best, but because the heavier tools carry a burden of justification that the lighter ones do not.

Regulatory cultures nonetheless differ. European regulation is dominated by a **risk-prevention principle**, which seeks to head off harms before they materialise, while regulation in the United States has more often accepted risk and compensated for loss only once a harm has actually occurred. For a company operating on both sides of the Atlantic, reconciling an ex-ante and an ex-post philosophy is a genuinely hard compliance problem, not a detail.

QUICK CHECK

Why does a liberal constitutional state, other things being equal, tend to prefer self-regulation or the lightest available regulatory instrument?

- Because self-regulation is always more effective than binding law
- **Because fundamental rights and proportionality require the least intrusive tool that still reaches the goal**
- Because the state lacks the legal power to regulate private firms
- Because self-regulation guarantees enforcement

1.4 The third way: regulated self-regulation

Between the two poles sits a hybrid that has become the workhorse of modern technology governance: **regulated self-regulation**. Here the state sets the goals and the procedural frame and then delegates the concrete rule-making to private actors, but it supervises the process and its results and keeps backup powers to intervene if the outcome does not match the goal. The precise definition comes from Schulz and Held (2002): self-regulation that is fitted into a state-set frame, or that takes place on a legal basis. The point is to capture the advantages of both poles, the expertise and speed of private actors and the accountability and enforceability of the state, while offsetting their weaknesses. It is a two-level system: a statutory frame above, a private regulatory instance below, with the state moderating, supervising and, in the last resort, sanctioning. The same idea travels under other names, co-regulation, enforced self-regulation, audited self-regulation, and it is exactly the structure the EU AI Act adopts when it invites codes of conduct and grants a conformity presumption to those who follow harmonised standards. Use the chooser to compare the three modes on the trade-offs a regulator actually weighs.

DEFINITION

Regulated self-regulation

Self-regulation fitted into a state-set frame (Schulz & Held 2002): the state fixes goals and procedure, delegates the detail to private bodies, supervises the outcome and retains backup powers. It combines private expertise with public accountability.

DEEP DIVE **Deep dive: when does regulated self-regulation actually work?**

Schulz and Held identify the conditions under which the hybrid succeeds. It works best where the subject matter is complex and fast-changing, so that the state suffers an information deficit that private actors can close; where creativity and initiative are wanted that command-and-control cannot compel; and where a direct legal command would simply bounce off an autonomous social system. It fails where the self-interest of the actors runs systematically against the regulatory goal, the classic case of a negative externality that the market will not price on its own. AI governance sits awkwardly across this line: the technical complexity argues strongly for delegation, yet the commercial incentive to move fast and capture markets can pull directly against the safety goals a code is meant to secure. That tension is the reason the AI Act frames the codes rather than simply trusting them.

02
KAPITEL

Internal codes of ethics

2.1 What a corporate code of ethics is, in legal terms

Against this map, an **internal code of ethics** locates itself precisely: it is an instrument of *explicit, organisation-internal self-regulation*. A firm writes down the values and rules it professes to follow, publishes them, and commits, on paper, to abide by them. In the vocabulary of the previous unit, a code of this kind is **soft law** of the softest sort. It is self-drafted, voluntary and, in the typical case, unaudited by any independent body and unenforceable against the firm by anyone outside it. A code can be revised or quietly abandoned by the same management that adopted it, and no authority will impose a penalty for the change. That does not make codes worthless, but it does mean their legal quality must not be overstated: they articulate expectations, they do not compel.

2.2 Two functions: compliance and public relations

Why, then, do firms write them? A code of ethics serves two functions that are worth holding apart. The first is a genuine **compliance and steering** function: a well-drafted code translates abstract values into concrete guidance for staff, coordinates behaviour across a large organisation, and can be woven into hiring, training and review. The second is a **public-relations and legitimation** function: the code signals responsibility to customers, regulators, investors and prospective employees, and it does so whether or not it changes anyone's conduct. These functions can coexist, but they can also come apart, and when a code performs only the second while claiming the first, it becomes precisely the **ethics washing** that Hagendorff diagnosed. The tell-tale signs are familiar: no independent audit, no published results, no grievance route for affected people, no participation by workers or user groups in drafting, and no consequence for breach. A code with all the rhetoric and none of the accountability is a legitimation device wearing the costume of a commitment.

DEFINITION

The two functions of a code of ethics

Compliance function: translating values into concrete internal guidance that actually steers staff behaviour. Public-relations function: signalling responsibility to outside audiences. Ethics washing occurs when only the second is performed while the first is merely claimed.

QUICK CHECK

Which feature most clearly distinguishes a genuine internal commitment from mere ethics washing?

A glossy publication and a memorable set of principles
An internal ethics board appointed and paid by the company
Independent audit with published results and an effective grievance route for affected persons
A public pledge signed by the chief executive

03
KAPITEL

A live specimen: the OpenAI Preparedness Framework

3.1 What the framework does

The clearest way to see these tensions is to examine a real corporate document. OpenAI's **Preparedness Framework (version 2, 2025)** is a self-regulatory instrument that governs how the company decides whether a frontier model is safe enough to deploy. It takes a **risk-based approach** focused on risks of "severe harm", and it sorts risks into two groups: **tracked categories**, currently biological and chemical capabilities, cybersecurity and AI self-improvement, and a looser set of research categories under watch. For a tracked capability, the framework runs an assessment, automated or a deeper manual "deep dive", against two thresholds, **High** and **Critical**, with escalating safeguards attached to each. Only once safeguards are implemented and tested does the framework permit a decision to deploy. As a piece of process design it is more concrete than most public codes, and it deserves credit for putting deployment behind explicit gates.

3.2 Reading it critically

Read against the matrix of values we assembled in the previous unit, however, the framework shows characteristic gaps. It foregrounds essentially a single value, **security**, yet its own risk categories implicate a far wider set: persuasion and model autonomy raise questions of **self-determination**, and the manipulation of information touches **freedom of opinion formation** and **equal access to communication**. The scope of the tracked categories is narrow, and the thresholds, definitions and criteria are presented with little methodological justification, no references, no account of how the numbers were derived, so that a reader cannot tell how the framework was built or by whom.

Most tellingly, the **governance structure** keeps decision authority inside the firm. A Safety Advisory Group offers expert advice and recommends mitigations "as targeted and non-disruptive as possible", but its members are appointed by OpenAI leadership; the CEO, or a designate, is the default decision-maker and may decide without the group; and the Board of Directors supervises leadership's own implementation of the framework. Every check on the process is internal to the very organisation the process is meant to constrain. This is self-regulation in its purest form, and it exhibits the structural weakness the whole unit has been circling: when the self-interest of the actor can diverge from the public goal, a self-appointed, self-supervised body is exactly the arrangement least able to hold the line.

04
KAPITEL

From principles to workflows: value-driven design

Between a code's abstract values and a working system lies a design process. Unit 4 introduced Virginia Dignum's (2019) **ART principles**, **Accountability** of the system, distributed **Responsibility** across all stakeholders, and **Transparency** of the system, together with the value-driven design method that carries values through norms into concrete functionalities rather than bolting ethics on at the end. The families of standard below are what happens when that design process is written down and made checkable by someone other than its author.

05
KAPITEL

Standards that operationalise ethics

Value-driven design is a method; standards are the shared, documented forms that let many organisations practise it consistently and let outsiders check that they did. This is where ethics finally meets the workflow, and three families of standard matter most for AI.

5.1 IEEE 7000 series

The IEEE developed a family of **process standards** aimed directly at building ethics into system design. The anchor is **IEEE 7000-2021**, a standard model process for addressing ethical concerns during system design, which operationalises value-driven design of the kind Dignum describes. Around it sit companions for specific concerns: **7001-2021** on the transparency of

autonomous systems, **7002-2022** on data-privacy process, **7005-2021** on transparent employer data governance, **7007-2021** an ontological standard for ethically driven robotics, and **7010-2020** on assessing the impact of autonomous systems on human well-being, alongside the child-focused age-appropriate-design framework in **2089-2021**. These are voluntary and, through the IEEE GET programme, freely accessible. Their distinctive contribution is process: they tell an organisation *how* to surface and address ethical concerns systematically, rather than prescribing a particular outcome.

5.2 NIST AI Risk Management Framework

On the American side, the **NIST AI Risk Management Framework** is a voluntary, freely available framework for identifying and managing the risks of AI systems across their lifecycle. It emerged in the policy climate around the US Executive Order 14110 on safe, secure and trustworthy AI, issued in October 2023, though that order was revoked in January 2025, which leaves the framework itself standing as a widely used voluntary reference even after its political scaffolding was removed. Like the IEEE process standards it prescribes no substantive result; it offers a structured, risk-based method that organisations can adopt of their own accord.

5.3 ISO/IEC 42001

The one instrument in this group that is **certifiable** is **ISO/IEC 42001:2023**, a management-system standard for artificial intelligence. It follows the familiar logic of ISO management-system standards such as those for quality or information security: an organisation builds a documented, auditable **AI management system** and can then be certified by an accredited third party as conforming to it. Its companion, **ISO/IEC 42005:2025**, addresses AI system impact assessment. Unlike the IEEE and NIST documents, these ISO standards are sold rather than given away, and certifiability is exactly what gives 42001 a different character: it converts self-regulation into something an outside party can verify, and it slots directly into the AI Act's compliance architecture, where quality management under the Regulation maps onto ISO/IEC 42001. Here the language of soft standards and hard law begins to merge.

With all three families now on the table, it helps to line them up side by side. Pick each standard in the comparator to see what kind of instrument it is, whether an outside party can certify it, and how it plugs into the AI Act's compliance architecture.

DEEP DIVE **Deep dive: why a harmonised standard is more than a technical document**

Under the EU AI Act, whoever develops a high-risk system in line with **harmonised standards**, those for which references have been published in the Official Journal, enjoys a **presumption of conformity** with the corresponding requirements, so that standardisation moves from a voluntary best practice into the core of the compliance system. That gives standards an unusual social role. A single harmonised standard is read by engineers as a technical specification, by regulators as a legal benchmark and by market actors as an economic fact, three worlds using one artefact for different purposes without needing to agree on its meaning. Objects with exactly this property, robust enough to hold a shared

identity yet flexible enough for local use, are what Star and Griesemer (1989) call **boundary objects**. Seen this way, the standards in this unit are not merely lists of requirements: they are the coordinating artefacts that let ethics, engineering and law act together, which is also why the fight over who writes them, and who can afford to read them, is a fight over power and not only over technique.

EXERCISE · TERMS

Among the AI standards, only **ISO/IEC 42001** is a certifiable management-system standard verified by an accredited third party, whereas **IEEE 7000** and the **NIST AI RMF** are voluntary process frameworks that prescribe method rather than a certified outcome.

QUICK CHECK

Which statement about the three standard families is correct?

All three can be certified by an accredited third party
ISO/IEC 42001 is a certifiable management-system standard, while IEEE 7000 and the NIST AI RMF are voluntary process frameworks
The NIST AI RMF is legally binding in the European Union
IEEE 7000 prescribes specific numerical risk thresholds for deployment

06

KAPITEL

From self-binding to binding law

This unit has traced a single arc. It began with a map of regulation graded by intensity and the reasons a liberal state leans towards the lightest tool, which explains the structural room left for firms to govern themselves. It then showed why pure self-regulation, whether a corporate code of ethics or the OpenAI Preparedness Framework, is strong on expertise and speed but weak exactly where self-interest and the public goal diverge, and how value-driven design and the IEEE, NIST and ISO standards try to turn principles into checkable process, with certifiability and harmonised standards forming a bridge into hard law. That bridge is where regulated self-regulation lives, and it leads straight into the next unit, which examines the hardest instrument of all: the EU AI Act, the binding European law that sets the frame these voluntary standards are increasingly built to satisfy.

07
KAPITEL

References

7.1 Literature

- 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
- 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
- Ogus, A. (2004): *Regulation: Legal Form and Economic Theory*. Hart Publishing, Oxford.
- Schulz, W. & Held, T. (2002): *Regulated Self-Regulation as a Form of Modern Government*. Hans-Bredow-Institut / University of Luton Press, Eastleigh.
- Star, S. L. & Griesemer, J. R. (1989): *Institutional Ecology, "Translations" and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39*. Social Studies of Science 19(3), 387-420. https://doi.org/10.1177/030631289019003001

7.2 Norms & Standards

- OpenAI (2025): *Preparedness Framework, Version 2*, last updated 15 April 2025. https://openai.com/index/updating-our-preparedness-framework/
- IEEE (2021): *IEEE 7000-2021, IEEE Standard Model Process for Addressing Ethical Concerns during System Design*. IEEE Standards Association, Piscataway. https://standards.ieee.org/ieee/7000/6781/
- IEEE (2021-2022): IEEE 7001-2021 (Transparency of Autonomous Systems), 7002-2022 (Data Privacy Process), 7005-2021 (Transparent Employer Data Governance), 7007-2021 (Ontological Standard for Ethically Driven Robotics and Automation Systems), 7010-2020 (Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being), 2089-2021 (Age Appropriate Digital Services Framework). IEEE Standards Association, Piscataway.
- NIST (2023): *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1. National Institute of Standards and Technology, Gaithersburg. https://doi.org/10.6028/NIST.AI.100-1
- ISO/IEC (2023): *ISO/IEC 42001:2023, Information technology - Artificial intelligence - Management system*. International Organization for Standardization, Geneva. https://www.iso.org/standard/81230.html
- ISO/IEC (2025): *ISO/IEC 42005:2025, Information technology - Artificial intelligence - AI system impact assessment*. International Organization for Standardization, Geneva.
- 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
- US Executive Order 14110 of 30 October 2023 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (revoked 20 January 2025).