thws Technical University of Applied Sciences Würzburg-Schweinfurt

FAKULTÄT INFORMATIK UND WIRTSCHAFTSINFORMATIK

LECTURE NOTES

Responsibility & the Responsibility Gap

Aristotle's conditions, the many-hands problem and value-driven design

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 saw that AI can act, and can even be described as an agent in an actor-network, without being a moral agent in the full sense. This unit takes the next step and asks who, then, must answer when an AI system causes harm. Responsibility is the hinge between the ethics of the earlier units and the enforceable duties of the legal units to come, because law can only assign liability once ethics has clarified who could plausibly be held to account in the first place. We begin with the classical conditions of responsibility, expose the distinctive ways in which AI strains them, and end with the constructive answer: building responsibility into the development process from the very start.

01
KAPITEL

What responsibility requires

1.1 The two conditions after Aristotle

Long before anyone imagined a learning machine, Aristotle set out the conditions under which a person can be held morally responsible for an action and its consequences. Two conditions must be met together. 1. The **control condition** is satisfied when the agent brings about the action, either by performing it directly or by exercising a sufficient degree of control over how it unfolds. 2. The **epistemic condition** is satisfied when the agent is aware, or could reasonably have been aware, of what they are doing and of the alternatives open to them.

Both conditions are tied to freedom of will and intent: the action must be *voluntary*, not forced and not carried out in self-inflicted ignorance. Aristotle already framed responsibility around two control questions, namely whether there was an alternative way of deciding or acting, and whether the agent was aware of that alternative, or at least could have been and chose to ignore it.

DEFINITION

The two conditions of responsibility

An agent is morally responsible for an action only if both the *control condition* (the agent causes it or sufficiently controls it) and the *epistemic condition* (the agent is or could be aware of what they are doing) are met, and the action is voluntary.

1.2 Why AI is not a full moral agent

Apply these conditions to an AI system and the result is clear. At the present state of technology it makes no sense to demand that an AI act voluntarily or free of ignorance, because an AI has no will of its own and no consciousness or awareness. It can *act*, in the sense of producing outputs that influence the world, and in that limited sense it has agency. It cannot, however, satisfy either Aristotelian condition as a moral matter: it neither controls its behaviour through free choice nor is aware, in any morally relevant way, of what it is doing. This is the point Floridi and Sanders

capture when they argue that artificial agents can be *accountable* for their behaviour, in the sense that their actions can be traced and attributed at a suitable level of abstraction, without therefore being morally *responsible*, which would presuppose the mental states that machines lack. The upshot is not that responsibility disappears, but that it must remain with the humans around the system.

QUICK CHECK

Why is an AI system not a full moral agent in the Aristotelian sense?
- Because it is not yet technically reliable enough
- **Because it lacks free will and awareness, so it satisfies neither the control nor the epistemic condition**
- Because it never makes mistakes
- Because it cannot act autonomously

02
KAPITEL

The problem: harm without an answerable agent

If AI is not a moral agent, and humans are, then the practical question becomes to whom, specifically, responsibility can be attributed. AI systems can produce great benefits, but they can also cause severe harm when errors, flaws and biases arise, and the honest answer to “who is responsible?” is rarely a single name. Two structural features make attribution hard. Many *hands* are involved, since designers, developers, vendors, deployers and users each touch only a fragment of the system. And many *things* are involved, since hardware, training data, models and interfaces interact in ways no participant fully surveys. The difficulty is not academic: it also shapes the law that regulates legal liability for AI, which we treat in a later unit. The ethical problems cluster into two families, control problems and knowledge problems.

2.1 The responsibility gap

Andreas Matthias gave the first family its name. Learning automata acquire their behaviour partly from data and experience after they leave the workshop, so their concrete conduct is neither fully foreseeable by the developer nor fully controllable by the operator. This opens a **responsibility gap**: a space in which harm is caused, yet no human satisfies the control and epistemic conditions well enough to be fairly held responsible. The developer could not foresee the specific behaviour, the operator cannot fully see through the system, and the user has delegated the decision. The gap sharpens wherever decisions fall on inhuman timescales, in self-driving cars, high-frequency trading or automated defence systems, because there the reassuring figure of the “human in

the loop" becomes a fiction: human supervision is meaningless when there is no real chance to intervene. Taken seriously, the responsibility gap even raises the question of whether it sets a principled limit to automation itself.

DEFINITION

Responsibility gap (Matthias 2004)

The space that opens when a learning system causes harm that the developer could not foresee and the operator cannot fully control, so that no human clearly satisfies the conditions of responsibility.

2.2 The problem of many hands

The second control problem is diffusion. In the development and deployment of AI, many organisations and far more individuals are typically involved, so it becomes genuinely unclear who should or must take responsibility. This **problem of many hands** does not abolish responsibility; it distributes it. But distributed responsibility only works under two demanding conditions. Ex ante, it requires a clear attribution of stakeholder roles before the system is built and run, so that everyone knows in advance which duties are theirs. Ex post, when something has gone wrong, it requires a deep technical, organisational and historical case analysis: who did what, when and where, and which components of the complex system were in fact causally relevant to the harm. Beyond individual and organisational responsibility, there is also a layer of collective responsibility, borne by organisations, states and cultures whose structures shape what the technology becomes.

The tool below turns this abstract analysis into a concrete judgement. Distribute responsibility across the parties in three scenarios and watch how quickly the responsibility gap opens, and where diffusion tempts you to leave harm unattributed.

2.3 The knowledge problems

The third family concerns not control but awareness, and so bears directly on Aristotle's epistemic condition. Developers and users are often unaware of the unintended consequences and the moral significance of an AI application, especially when a system is misused or carried into a context for which it was never built, as when a face-recognition tool designed for one purpose is redeployed for public-security surveillance. This raises a straightforward question: how can developers and users be *enabled* to acquire the relevant knowledge? The answer points to transparency and literacy measures. The problem is compounded by the **black-box** character of most modern AI systems, whose lack of technical transparency and functional explainability (how the system actually produces a given output) can lead to overreliance and, worse, to a quiet offloading of responsibility onto a machine that cannot carry it.

CASE STUDY

Case: the repurposed face-recognition model

A start-up releases a face-recognition model trained and documented for photo organisation in consumer apps. A city police department procures it and uses it to identify suspects from CCTV footage. In this new setting the model misidentifies people from certain demographic groups at much higher rates, leading to wrongful stops. Where does responsibility lie, and which condition is chiefly at stake?
Solution. The harm here flows less from a loss of control than from a failure of knowledge, so the **epistemic condition** is decisive. The developer did not foresee the security use, and the deployer did not understand the model's limits outside its documented purpose. Crucially, unforeseeability does not dissolve responsibility; it relocates it as a duty to inquire. The developer bears responsibility for stating the intended purpose and known limits clearly, and for not marketing the tool for uses it cannot support. The deployer, who chose to run a consumer tool in a high-stakes context, bears responsibility for the due diligence it omitted: validating error rates on its own population before relying on the output. The "many things" dimension is visible too, since biased training data and CCTV image quality interact. The right response is not to blame the system but to reconstruct, ex post, who could and should have known what, and to fix the ex-ante role assignment that let a consumer tool drift into policing.

03
KAPITEL

From control to answerability: transparency, explainability, accountability

3.1 A relational turn

The classical account we have used so far focuses on the moral *agent* and on their control and knowledge. Coeckelbergh (2020) proposes a productive shift of focus toward the relation between the agent and the moral *patient*, the person affected by the action. On this relational view, transparency and explainability are not merely technical desiderata but themselves moral obligations, understood as the **answerability** of moral agents to moral patients. The patient has a moral right to demand an explanation from the responsible agent, and it follows that technical transparency and explainability of AI systems are morally required precisely in the service of the human agents who owe answers to the patients affected by their systems. This is where **Responsible Research and Innovation** enters: if agents owe answers, they must organise their development so that answers can in fact be given. Explainable AI (XAI) is the technical research programme that tries to supply the tools for exactly this.

3.2 Three concepts that are easily confused

The vocabulary of the field packs three distinct ideas into words that are often used interchangeably. **Keeping them apart is one of the most useful things this unit can offer.**

Accountability arises when a task is delegated. The operator, the "agent" in the principal-agent sense, must be able to explain to the delegating "principal" how and why the task was carried out in a particular way; responsibility towards third parties, however, remains with the principal who delegated. This logic transfers directly to digital systems: an AI to which a task is delegated should be *accountable*, able to report how it did the task, while the human principal remains responsible. In Virginia Dignum's framing, accountability is essentially backward-looking, the ex-post capacity to explain what was done, whereas responsibility is forward-looking, the ex-ante duty to answer for actions.

Explainability is the ability of those responsible to explain their decisions and actions to those affected. It is patient-facing: it answers the moral patient's demand for reasons. Explainable AI is the effort to build technical instruments that make such explanation possible.

Transparency is the ability of those responsible to understand and describe the technical inter-relationships of the system in the first place; it is a precondition of meaningful control and of any honest explanation. For Dignum, transparency reaches far beyond publishing source code, which she calls a red herring, since code is uninterpretable to most people and entangled with intellectual property. Real transparency spans four dimensions: openness about the *data*, the *design process*, the *algorithms* and their optimisation goals, and the *actors and stakeholders* involved. Transparency thus feeds control, explainability serves answerability, and accountability structures the delegation between them.

EXERCISE · TERMS

When a principal delegates a task, the agent must be able to report how it was performed, which is **accountability**, while responsibility to third parties stays with the **principal**. The duty to give reasons to those affected is **explainability**, and the capacity to understand the system's technical workings in the first place is **transparency**.

QUICK CHECK

Which concept denotes the ability of a responsible party to explain a decision to the person affected by it?

Transparency
Explainability
Accountability
Robustness

DEEP DIVE **Deep dive: does explainability offload or anchor responsibility?**

There is a subtle danger in the enthusiasm for XAI. If an organisation treats an explanation produced by the system as if it settled the matter, explanation can become a new route for offloading responsibility: "the model explained itself, so we relied on it." The relational account guards against this. Explainability is owed by the *human* agents to the *human* patients; the technical tool merely helps those agents discharge a duty that remains theirs. A

good explanation should therefore make a human more answerable, not less, by surfacing the assumptions, data and limits for which a person can then be held to account. Used this way, XAI anchors responsibility in the principal rather than dissolving it into the machine.

04
KAPITEL

Why ethics must enter the development process

4.1 The Collingridge dilemma

If responsibility is hard to assign after harm, the natural move is to intervene earlier. But *how much* earlier is itself a dilemma, sharply stated by David Collingridge in 1980. Early in a technology's life, **design flexibility** is high: the system can still be shaped, redirected or abandoned at modest cost. Yet at that same early stage our **knowledge of the technology's effects** on society is low, because the effects have not yet materialised. Later, once a technology is embedded in social practice and infrastructure, we finally understand its effects, but by then flexibility has collapsed: we depend on the technology's functions, and it is locked into existing processes. The two curves run in opposite directions, so the moment when we know enough to steer wisely is the moment when steering has become most costly. This is precisely why purely *ex post* regulation is structurally too late, and why ethical assessment must begin during development rather than after deployment.

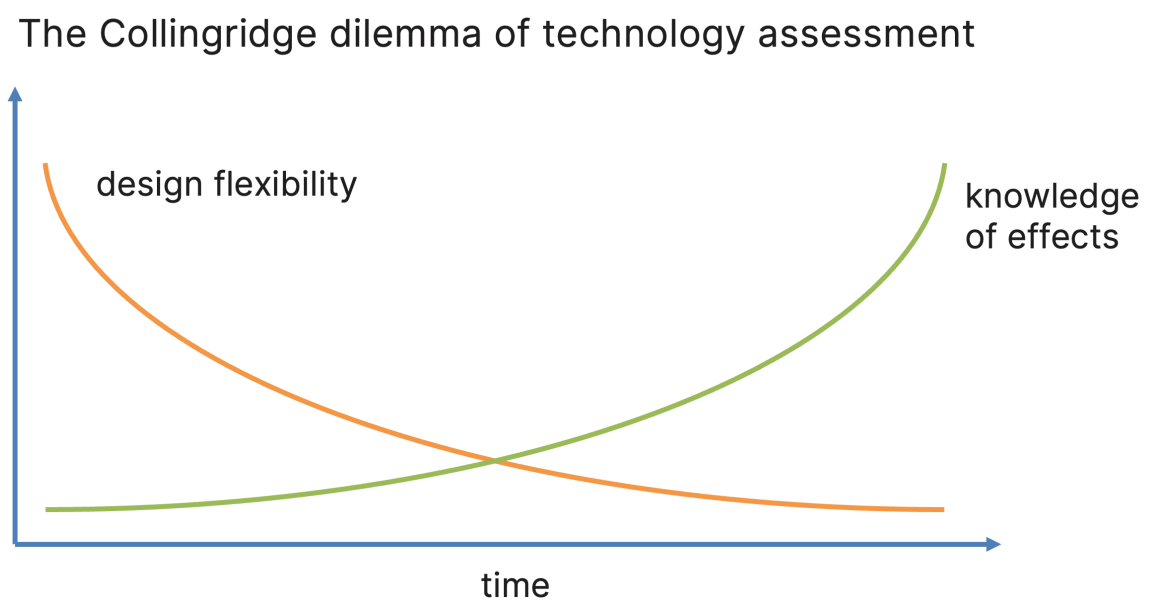


Figure 1: The Collingridge Dilemma: Design flexibility decreases over time, while knowledge of the effects increases—the two curves move in opposite directions and with increasing steepness, so that the dilemma does not come to a head until late in the development process.

4.2 Responsible Research and Innovation

The conclusion Collingridge forces is constructive. Since neither pure foresight nor pure hindsight suffices, the response must be procedural: establish binding ethical guidelines and assessment procedures for all the stakeholders, that is, all the moral agents, involved in developing and deploying AI, and do so throughout the process rather than at its end. This is the programme of **Responsible Research and Innovation (RRI)** in and of AI, an ongoing, participatory practice of anticipating effects, reflecting on values and adjusting the design while adjustment is still cheap. RRI operationalises the relational demand for answerability at the level of the whole innovation system: it distributes the many hands deliberately, and it manufactures, as early as possible, the knowledge that the Collingridge dilemma otherwise withholds.

QUICK CHECK

What does the Collingridge dilemma imply for the governance of AI?

Regulation should wait until a technology's effects are fully understood.

Purely ex-post regulation comes too late, so ethical assessment must accompany development from the start.

Design flexibility and knowledge of effects both rise together over time.

Early-stage technologies are impossible to change.

05
KAPITEL

Dignum's ART principles and value-driven design

5.1 The ART principles

Virginia Dignum (2019) brings these threads into a workable framework. Her starting point is that AI systems must be treated, even at early stages of development, as **sociotechnical systems** whose relevant stakeholders and societal actors should be integrated as far as possible into an RRI process of *open development*. Development and deployment should then follow the **ART principles**: **A**ccountability of the AI system, distributed **R**esponsibility of all the stakeholders involved (secured through an explicit distribution agreement), and **T**ransparency of the AI system. The three are not arbitrary; Dignum pairs each with one of the defining characteristics of AI, so that accountability answers to autonomy, responsibility to interaction, and transparency to adaptivity. Read this way, ART is a set of counterweights, one for each property that makes AI ethically demanding.

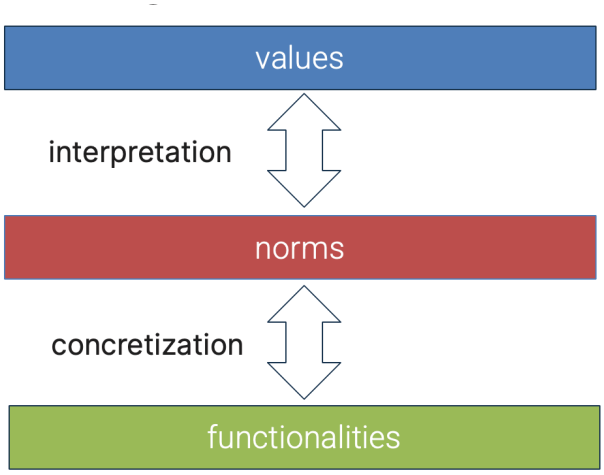
DEFINITION

The ART principles (Dignum 2019)

Accountability (the system can report how and why it acted), distributed *Responsibility* (all human stakeholders answer for their roles under an explicit distribution agreement), and *Transparency* (openness about data, design, algorithms and actors).

5.2 From values to functionalities

ART would remain a slogan without a method to carry values into code, and this is what **value-driven design** (also called design for values) provides. Its core problem is that values are highly abstract and do not translate directly into system requirements; when the translation is left implicit, ethically loaded design choices become untraceable. Dignum's model therefore introduces two intermediate steps on a bidirectional ladder. **Values**, such as fairness, are interpreted into **norms**, such as the concrete choice between "equal access" and "equal opportunity", which are in turn concretised into **functionalities**, the actual features of the system. Reading upward, functionalities *count as* the realisation of norms in a given context, and norms *interpret* values. The point of making these links formal is traceability: when a value conflict surfaces, designers can follow the chain from a functionality back to the value it was meant to serve, and adjust deliberately rather than by accident.



Cf. Dignum 2019: p. 63

Figure 2: Dignum's value-driven design ladder: values are interpreted into norms, which are in turn concretised into functionalities - a bidirectional chain that keeps the design traceable back to the values it serves (own illustration, based on Dignum 2019: p. 63).

The method is embedded in a repeatable **process structure** that turns the abstract ladder into concrete project steps. The whole is a continuous loop rather than a one-off checklist, so that the system can be readjusted as values and perceptions shift over time.

1. Identify the relevant stakeholders.
2. Elicit the values and requirements of all parties, for instance in workshops. -> Aggregate the values and their competing interpretations.
3. Derive the values into norms and functionalities.
4. Record the formal links between values, norms and system functionalities, so the system stays adjustable.
5. Support everyone in the R&D process in choosing components in light of the underlying societal and ethical commitments.

5.3 Dignum's process structure

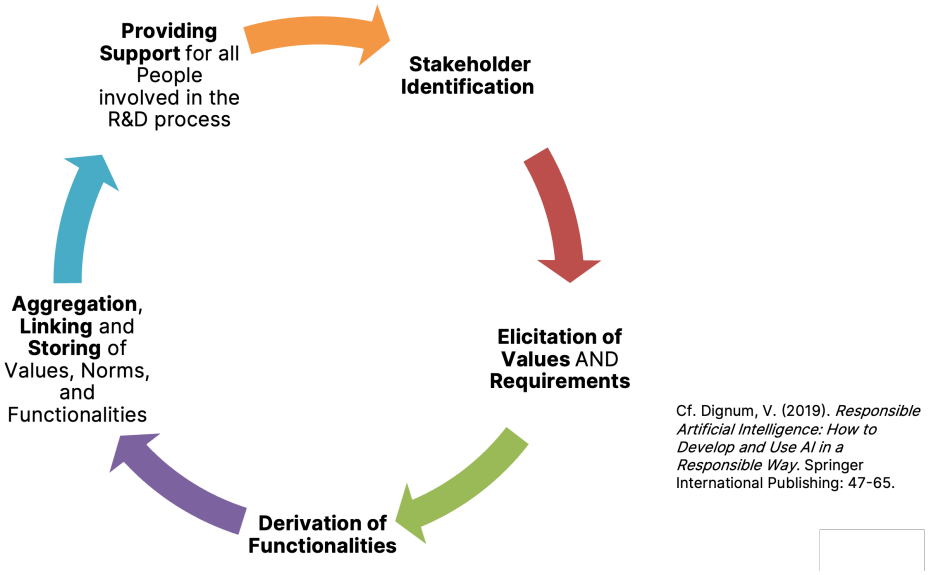


Figure 3: Dignum's process structure as a continuous loop: stakeholder identification feeds the elicitation of values and requirements, which are derived into functionalities, then aggregated, linked and stored as values, norms and functionalities, and finally used to support everyone involved in the R&D process, whose work in turn refines stakeholder identification (source: Dignum 2019).

5.4 From ART to codified standards

Value-driven design is no longer only an academic proposal; it has been translated into industry standards that turn abstract principles into auditable process requirements.

The IEEE 7000 series is the clearest example. IEEE 7000-2021 defines a model process for addressing ethical concerns during system design, IEEE 7001 addresses the transparency of autonomous systems, IEEE 7002 specifies a data-privacy process, and IEEE 2089 sets out age-appropriate digital services. These standards matter because they convert the "ought" of ethics into checkable steps that an organisation can adopt, audit and be held to, forming a bridge from the voluntary ethics of this unit to the binding compliance duties of the regulatory units that follow.

06
KAPITEL

From responsibility to fairness

We have moved from a classical account of responsibility, through the distinctive gaps that AI opens in it, to a constructive programme for closing those gaps by design. Responsibility, we saw, cannot rest with the machine; it must be distributed among human agents, made answerable through transparency and explainability, and built in from the start through value-driven design and the ART principles. Yet one value kept surfacing as the hardest to specify and the easiest

to violate, namely fairness. The next unit takes it up directly, asking how bias and discrimination enter AI systems and how, if at all, they can be measured and constrained.

07
KAPITEL

References

7.1 Literature

- Coeckelbergh, M. (2020): Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics* 26 (4), pp. 2051-2068. https://doi.org/10.1007/s11948-019-00146-8
- Collingridge, D. (1980): *The Social Control of Technology*. Frances Pinter, London.
- Dignum, V. (2019): *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (Artificial Intelligence: Foundations, Theory, and Algorithms). Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-30371-6
- Floridi, L. & Sanders, J. W. (2004): On the Morality of Artificial Agents. *Minds and Machines* 14 (3), pp. 349-379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d
- 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

7.2 Norms & Standards

- IEEE 7000-2021: Model Process for Addressing Ethical Concerns during System Design.
- IEEE 7001-2021: Transparency of Autonomous Systems.
- IEEE 7002-2022: Data Privacy Process.
- IEEE 2089-2021: Age Appropriate Digital Services Framework.