



Philosophical Foundations of AI Certification

Victor Weisbrod

Imprint

Author

Victor Weisbrod

Center for Science and Thought

Universität Bonn

Regina-Pacis-Weg 3

53113 Bonn

Germany

<https://www.cst.uni-bonn.de/en>

The Center for Science and Thought (CST) at the University of Bonn, Germany, is a radically interdisciplinary platform designed to address urgent questions that arise at the intersection of philosophy and various natural sciences.

“Zertifizierte KI” (Certified AI) is a KI.NRW-flagship project funded by the Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen (Ministry for Economic Affairs, Industry, Climate Action and Energy of the State of North Rhine-Westphalia).

Cover photo

‘bottom view of bridge during day’, by valor kopeny on Unsplash

<https://unsplash.com/photos/bottom-view-of-bridge-during-day-Q6NCx-h7KpY>

Layout and design

Victor Weisbrod

Bonn, October 2026

Contents

Preliminary Remarks	3
1 Introduction	4
2 The Dominant Paradigm: Principles	5
2.1 The General Idea and How It Became Dominant	5
2.2 Formulation Assumption	6
2.2.1 Principle Convergence and Variance	6
2.2.2 Method of Stakeholder-Informed Reflective Equilibrium	7
2.2.3 Three Worries	8
2.3 Translation Assumption	8
2.3.1 The Principle-Practice Gap	8
2.3.2 Frameworks of Principle-To-Practice Translation	9
2.3.3 Four Worries	10
2.4 Arbitration Assumption	11
2.4.1 Value Monism and Value Pluralism	11
2.4.2 Proposals for Principle Conflict Arbitration	12
2.5 Isolation Assumption	13
2.5.1 Individual Artefact as Unit of Assessment	13
2.5.2 Two Worries	14
3 An Alternative Paradigm: Systems	16
3.1 The General Idea of Systemic AI Certification	16
3.2 Holism Assumption	17
3.2.1 Complex System as Unit of Assessment	17
3.2.2 Implicit Normativity of Complex Systems	18
3.2.3 Advantage and Two Worries	19
3.3 Mapping Assumption	20
3.3.1 Systems Cartography 1: Elements, Relations, Dynamics, and Boundaries	20
3.3.2 Systems Cartography 2: Explicating Implicit Norms	20
3.3.3 Advantage and Two Worries	21
3.4 Diagnosis Assumption	22
3.4.1 Life-Cycle Assessment 1: Overarching Framework	22
3.4.2 Life-Cycle Assessment 2: Sphere-Specific Diagnostics	24
3.4.3 Advantage and Two Worries	24
3.5 Evaluation Assumption	25
3.5.1 Method of Immanent-Critical Evaluation	25
3.5.2 Advantage and Two Worries	26
4 Conclusion and Outlook	28
5 References	30

Preliminary Remarks

Executive summary

This white paper conducts a systematic inquiry into the philosophical foundations of AI certification, that is, into the *basic assumptions of the institutionalised endeavour to assess AI applications as ethically good or bad*. It identifies and examines two paradigms of AI governance: the dominant *principles paradigm* and the alternative *systems paradigm*. By clarifying and problematising the foundational assumptions of both paradigms, the white paper elucidates their respective strengths and weakness. The principles paradigm provides an administratively legible, legally predictable, and standardised framework for product-level auditing. However, it encounters significant challenges regarding underdetermination, value arbitration, narrow artefact boundaries, and ethics washing. The systems paradigm approaches AI applications in a manner that aims to capture these as embedded in complex ecological, social, and economic systems with inherent standards of good functioning. In turn, it faces significant hurdles concerning system unboundedness, data opacity, normative projection, and temporal bottlenecks. Ultimately, by explicating the implicit assumptions and limitations of both paradigms, the white paper aims to deepen our understanding of AI certification as a complex governance practice, and to provide foundational insights for future research, standard-setting, and policymaking in this field.

Structure of white paper

The white paper is divided into four chapters. Chapter 1 introduces the societal and regulatory landscape of AI certification, the notion of inquiring into its foundational assumptions, and the two paradigms of AI certification. Chapter 2 elaborates and critically evaluates the dominant principles paradigm of AI governance. The chapter is

divided into five sections that spell out and problematise the general idea (sec. 2.1) and the four basic assumptions of the paradigm: the formulation (sec. 2.2), translation (sec. 2.3), arbitration (sec. 2.4), and isolation assumptions (sec. 2.5). Chapter 3 provides a similar treatment of the systems paradigm of AI governance. This chapter is also divided into five sections that clarify and problematise the general idea of systemic AI certification (sec. 3.1) and the paradigm's four basic assumptions: the holism (sec. 3.2), mapping (sec. 3.3), diagnosis (sec. 3.4), and evaluation assumptions (sec. 3.5). Finally, chapter 4 concludes by reflecting on and synthesising the findings about the two paradigms, and by indicating future directions for research, standard-setting, and policymaking in AI governance.

Context

This publication is the final white paper produced within the philosophical sub-project of the KI.NRW flagship project *Certified AI*, funded by the Ministry of Economic Affairs, Industry, Climate Action and Energy of the German State of North Rhine-Westphalia, and based at the Center for Science and Thought at the University of Bonn, from 2021 to 2026. It builds directly upon the foundational research presented in the preceding white papers of the project: “Trustworthy Use of Artificial Intelligence”,¹ “Catalog of General Ethical Requirements for AI Certification”,² and “Catalog of Field-Specific Ethical Requirements for AI Certification”.³ Having taken over the direction of the philosophical sub-project from Nicholas Kluge Corrêa and Julia Maria Mönig in April 2026, I have dedicated this final white paper to a critical meta-reflection on the preceding philosophical work of the project, as well as on the broader field of AI ethics and certification.

¹ Cremers et al. 2019

² Corrêa and Mönig 2024. See also Corrêa and Mönig 2026b

³ Corrêa and Mönig 2026a

I Introduction

Artificial intelligence (AI) has transitioned from speculative science fiction into a ubiquitous feature of our daily lives. Driven by rapid breakthroughs in deep learning and the proliferation of large-scale foundation models, AI applications are now deeply embedded in many critical social domains. They influence high-stakes outcomes such as clinical diagnostics in healthcare,⁴ risk assessment in criminal justice,⁵ recruitment and worker management in human resources,⁶ creditworthiness evaluation in financial services,⁷ and performance of students.⁸ Given their profound impacts on social relations, individual lives, as well as the environment, AI applications raise pressing concerns regarding how to progress technologically while remaining true to our core ethical values.

To address these challenges, society has turned to AI certification: the institutionalised endeavour to assess whether AI applications live up to a set of standards for lawful and ethical use. In essence, AI certification represents the institutionalisation of AI ethics. It is the institutionalised endeavour to systematically distinguish between *good* and *bad*, *ethical* and *unethical* AI applications.⁹ This endeavour is currently being pursued by national, transnational, and international bodies alike. Foremost among these is the EU's *Artificial Intelligence Act* (2024), which establishes an overarching, risk-based framework for conformity assessment and certification. Similar standard-setting and governance initiatives have been advanced globally, including the IEEE's *Ethically Aligned Design* (2019), the OECD's *AI Principles* (2019/2024) and *Tools for Trustworthy AI* (2021), UNESCO's *Recommendation on the Ethics of Artificial Intelligence* (2021), and the WHO's *Ethics and Governance of Artificial Intelligence for Health* (2021). On a more local scale, the KI.NRW-flagship project *Certified AI*, funded by the German State of North Rhine-Westphalia, exemplifies this institutionalised, transdisciplinary endeavour to develop concrete audit catalogues.¹⁰

When we ask about the *philosophical foundations* of AI certification, we are asking about the basic assumptions of this institutionalised endeavour to assess AI applications as ethically good or bad. In other words, we are asking about the basic assumptions that enable us to systematically identify, categorise, and regulate AI applications in terms of ethical goodness or badness. The aim of this white paper is to explicate these foundations of the two major paradigms of AI certification:

Principles paradigm: This is the dominant paradigm in AI ethics and certification. According to this paradigm, ethical governance of AI applications is conceived as a process of formulating a set of abstract ethical principles (such as 'fairness', 'privacy', and 'autonomy') and subsequently translating them into technical criteria against which AI applications can be assessed in practice.

Systems paradigm: This is an alternative paradigm, inspired by recent developments in AI ethics, social philosophy, philosophy of technology, and complex systems theory. It shifts the focus of evaluation from isolated technological artefacts to the broader eco-socio-economic systems in which they are embedded. Instead of applying abstract principles in a top-down manner, it seeks to evaluate the impact of AI applications on the functioning of these systems in light of concrete standards already operative in them.

By clarifying and problematising the basic assumptions underlying these two paradigms, the white paper seeks to elucidate the philosophical foundations of the institutional endeavour of AI certification. The aim is to uncover the deep commitments, but also the weak points of this endeavour that political and economic actors may seek to exploit, thereby contributing to the development of more robust, just, and sustainable AI governance frameworks.

⁴ Ashraf et al. 2025; Bajwa et al. 2021; Fahim et al. 2025

⁵ Farber 2026; McGregor Richmond and Primeau 2026; Zuhair et al. 2026

⁶ Madanchian et al. 2023; Nguyen et al. 2026; Pranavi et al. 2026; Úbeda-García et al. 2025

⁷ Ayari et al. 2026; Genovesi et al. 2024; Goyal et al. 2025

⁸ Adewale et al. 2024; Dong et al. 2025; Vieriu and Petrea 2025

⁹ Brogle et al. 2025; Corrêa and Mönig 2026b; Cremers et al. 2019; Genovesi and Mönig 2022

¹⁰ See Cremers et al. 2019; Corrêa and Mönig 2024; Corrêa and Mönig 2026a; Corrêa and Mönig 2026b

2 The Dominant Paradigm: Principles

2.1 The General Idea and How It Became Dominant

At its core, the principles paradigm conceives of the ethical governance of technologies as a top-down, translational procedure.¹¹ The general idea is straightforward: a multi-stakeholder body, academic cohort, or regulatory authority first formulates a set of abstract ethical principles or values, such as ‘fairness’, ‘privacy’, and ‘accountability’. These principles, often functioning initially as non-binding ‘soft law’ or voluntary guidelines,¹² are then unpacked into concrete technical criteria in audit catalogues. Finally, there is the auditing procedure itself, where an AI application is assessed against the standardised requirements, and if it conforms, it is awarded an ethical certification.

This principles-based understanding of the governance of technologies is the dominant paradigm in AI ethics and certification today. Crucially, it rests on four fundamental assumptions:

Formulation assumption: the general ethical principles can be readily formulated.

Translation assumption: the ethical principles can be successfully translated into technical criteria for assessing AI applications in practice.

Arbitration assumption: conflicts between ethical principles can be rationally arbitrated.

Isolation assumption: an AI application can be ethically assessed in relative isolation from the broader ecological, social, and economic system as part of which it is developed and deployed.

Before turning to spell out and problematise these basic assumptions in more detail, it is worth noting that the dominance of the principles paradigm is not a neutral result

or an historical accident. Rather, its rise is a consequence of powerful philosophical, political, and economic factors:

a) Philosophical factors: The principles paradigm aligns seamlessly with the dominant Anglo-American tradition of analytic philosophy, which, in the practical domain, is overwhelmingly concerned with formulating, systematising, and balancing normative principles.¹³ Furthermore, AI ethics did not emerge in a vacuum, but rather directly inherited key commitments of pre-existing domains of applied ethics, most notably bioethics.¹⁴ The highly influential pluralistic principlism developed by Beauchamp and Childress—which organises moral reasoning around four *prima facie* principles: ‘autonomy’, ‘beneficence’, ‘non-maleficence’, and ‘justice’—served as the direct blueprint for much of AI ethics.¹⁵ Consequently, there is a large conceptual overlap between the principles of bioethics and the core requirements of ethical AI.¹⁶

b) Political factors: For policymakers and state regulatory institutions, the principles paradigm presents an attractive governance framework. Rapid technological breakthroughs often generate profound epistemic and social side effects that outpace the structural timelines of traditional lawmaking.¹⁷ In this context of regulatory lag, a principled approach offers policymakers and regulators a way to react quickly and, at least on paper, effectively to the challenges of AI proliferation.¹⁸ It offers both a vocabulary for making visible, politically defensible commitments to regulating AI applications in the name of established ethical values, and clearly implementable standards that can easily integrate into existing auditing frameworks.

c) Economic factors: Private corporations and AI developers have also embraced the principled approach with enthusiasm. Because ethical principles are formulated in broad, abstract terms, they deliver standards that are

¹¹ Bolte et al. 2022; Corrêa et al. 2024; Max 2026. See also the various contributions in Floridi 2021

¹² OECD 2019/2024; Shuju and Aifen 2023; UNESCO 2021; Waelen 2022

¹³ See Harris and Dubljević 2025. See also Coeckelbergh 2020

¹⁴ Floridi and Cows 2019; Harris and Dubljević 2025; Mittelstadt 2019; Waelen 2022

¹⁵ Beauchamp and Childress 2019

¹⁶ See for instance Floridi and Cows 2019

¹⁷ Sætra and Rogerson 2026; Vinuesa et al. 2020. See also the contributions in Mazzi and Floridi 2023

¹⁸ Larsson 2021; Max 2026

highly flexible.¹⁹ This abstract nature works remarkably well within a capitalist market economy.²⁰ It allows companies to publicly commit to 'trustworthy AI' to cultivate brand reputation and gain consumer trust, while retaining maximum operational flexibility. Furthermore, the paradigm synergises well with a liberal political order: standardising ethics into flexible benchmarks permits market-led innovation to proceed rapidly, allowing businesses to treat ethical compliance as a competitive advantage rather than an obstacle.

In sum, the philosophical lineage of Anglo-American analytic philosophy and applied ethics, the political demand for immediate regulatory tools, and the economic incentives of corporate self-regulation have converged to make the principled approach the dominant paradigm in AI governance today.

2.2 Formulation Assumption

2.2.1 Principle Convergence and Variance

The first basic assumption of the principles paradigm is that the general ethical principles by means of which to assess AI applications can be readily formulated. On the face of it, this assumption seems entirely warranted. When reviewing the global landscape of AI ethics, we see a widespread tendency of convergence upon a set of core ethical principles. Meta-analyses of global guidelines show that various private, corporate, state, and intergovernmental bodies repeatedly endorse a relatively uniform set of principles.²¹ Again, this convergence is not a historical coincidence. Rather, it represents a conscious, multi-stakeholder endeavour to unify the normative vocabulary needed to assess and regulate AI applications, known as the *standardisation of AI ethics*.²²

The core principles of ethical AI are almost always modelled to correspond to, or at least include, the classical four principles of bioethics: 'autonomy', 'non-maleficence',

'beneficence', and 'justice'.²³ This structural parallel is explicitly reflected in several prominent AI governance frameworks, such as Floridi and Cowls' *unified framework* (2019), the EU's *Ethics Guidelines for Trustworthy AI* (2019), UNESCO's *Recommendation on the Ethics of Artificial Intelligence* (2021), Microsoft's *Responsible AI Standard* (2022), and Corrêa and Mönig's *Catalog of General Ethical Requirements for AI Certification* (2024). Synthesising and simplifying this sprawling landscape of governance frameworks, we find that twelve principles of ethical AI applications are most frequently proposed in the literature:

Autonomy	Accountability	Robustness
Non-maleficence	Safety	Sustainability
Beneficence	Fairness	Transparency
Justice	Privacy	Truthfulness

Despite this general convergence, there is still a considerable variance in the principles proposed in the various catalogues presented by researchers, inter- and non-governmental organisations, and corporations. While certain principles like 'non-maleficence' and 'transparency' are almost universally cited, others such as 'sustainability' and 'truthfulness' are significantly underrepresented, and others still only occur in singular catalogues, like 'awareness and literacy' in UNESCO's *Recommendation*.²⁴ Furthermore, there is a considerable variance in how (some of) these principles are defined, a variance, so one may suspect, that masks deep, unresolved normative differences over what is actually required for an AI application to count as fair or sustainable.²⁵

This persistent variance in both the catalogues and definitions of ethical AI principles raises a fundamental problem for the principled approach to AI ethics: how do we arbitrate between these different catalogues and definitions? That is, how do we decide which principles are to be included among the requirements of ethical AI applications, and how are we to define them? To establish an ethically and legally sound certification scheme, we must

¹⁹ Crawford 2021; Ong and Findlay 2023

²⁰ Brevini 2023; Crawford 2021; Dauvergne 2020; Orru 2026

²¹ Fjeld et al. 2020; Jobin et al. 2019; Corrêa et al. 2023

²² See Corrêa and Mönig 2026b; Lewis et al. 2020

²³ Beauchamp and Childress 2019

²⁴ See Hagendorff 2020; Jobin et al. (2019); Corrêa et al. (2023); Corrêa and Mönig 2024

²⁵ Max 2026; Mittelstadt 2019

be able to justify why we have included certain principles in our catalogue and why not others. However, this presupposes that we possess a reliable *method* for correctly formulating these principles, a methodological hurdle that sits at the very heart of the principles paradigm.

2.2.2 Method of Stakeholder-Informed Reflective Equilibrium

A striking feature of AI ethics is the relative silence about its own methodological foundations. While hundreds of policy guidelines, corporate charters, and international recommendations confidently put forward lists of ethical AI principles, they rarely mention, let alone explicate, the method(s) used to formulate these principles. It is often treated as a self-evident fact that these are indeed the correct ethical principles to govern AI technologies. However, if the principled approach to AI certification is to be a robust, legally and philosophically defensible enterprise, its practitioners must be following some method for formulating the principles of ethical AI use in a reliable and justifiable manner.

Upon closer examination, it becomes clear that practitioners operating within the principles paradigm tend to use a form of the method of *reflective equilibrium* that is informed by *stakeholder deliberation procedures*.²⁶ Reflective equilibrium is the dominant method in moral and political philosophy in the Anglo-American analytic tradition. It was first articulated by John Rawls in *A Theory of Justice* (1971/1999), with subsequent proponents declaring it to be “the only rational game in town for the moral theorist” and “the only defensible method”.²⁷ The method takes the form of a *process of mutual adjustment of particular beliefs and general principles* in light of relevant theories and empirical evidence. We begin with our “considered judgements” or well-informed, non-distorted beliefs about specific cases, for example, the belief that racially biased sentencing algorithms are unjust, or the belief that unreliable chat bots are harmful. We then attempt to

formulate general principles that can accommodate, explain, and extend these specific beliefs, ideally done by reflecting on a wide range of theories and arguments. Crucially, this process is dynamic and iterative: if a general principle we formulate conflicts with a deeply held belief, or if a specific belief clashes with our principles, we do not simply discard one or the other. Instead, we work back and forth, revising principles and beliefs until we reach a state of equilibrium in which they agree with each other.²⁸

What is distinctive of the application of reflective equilibrium in the context of AI ethics is that it begins, not so much from individually held beliefs, but rather from the results of stakeholder deliberations.²⁹ To arrive at the correct ethical principles of AI applications requires input from multiple stakeholders, including software engineers, corporate representatives, academic researchers, legal experts, and civil society groups. Through structured deliberative processes, such as workshops and public consultations, these diverse actors articulate well-informed beliefs about the concrete benefits, burdens, risks, and opportunities associated with the deployment of AI applications. It is based on these beliefs that ethicists may then set out to formulate a set of principles of ethical AI.

A paradigmatic example of this method of stakeholder-informed reflective equilibrium is the European Commission’s High-Level Expert Group on Artificial Intelligence (AI HLEG).³⁰ The AI HLEG convened 52 stakeholders from across industry, academia, and civil society to engage in an extensive social and ethical dialogue. On the basis of the well-informed, yet multifarious and divergent, beliefs articulated during this stakeholder deliberation, ethicists subsequently set out to formulate general principles through the aforementioned process of mutual adjustment towards equilibrium. The HLEG thus combined concrete stakeholder experience and general ethical expertise to formulate its widely cited seven key requirements for ethical AI.³¹

²⁶ This at least seems to be the most charitable interpretation of the method presupposed in the literature reviewed in this white paper. For more on the notion of ‘charitable interpretation’, see Davidson 2001, ch. 2; ch. 9; ch. 13; Gauker 1986.

²⁷ DePaul 1993, p. 6; Scanlon 2003, p. 149

²⁸ For more on the method of reflective equilibrium, see Cath 2016; Rawls 1999, pp. 42-45; Rawls 2003, pp. 29-32; Scanlon 2003; Scanlon 2014, pp. 69-104; Schmidt 2024

²⁹ Corrêa and Mönig 2026a; Genovesi and Mönig 2022; Genovesi et al. 2024; Maroccia 2026; Mensch et al. 2026; Redin and Ames 2026

³⁰ Bolte et al. 2022; Floridi and Cowlis 2019; Genovesi and Mönig 2022; Kindlydi and Cabral 2021

³¹ EU AI HLEG 2019

2.2.3 Three Worries

Although the method of stakeholder-informed reflective equilibrium is operationally convenient and rationally justifiable, it faces at least three serious worries:

1) **Mere rationalisation:** Because reflective equilibrium sets out to bring pre-existing beliefs into agreement with subsequently formulated principles, it can easily degenerate into a conservative exercise of self-justification. Rather than discovering objective standards of ethical AI applications, the method risks merely rationalising the contingent, historically situated beliefs that the participating stakeholders already hold.³²

2) **Empty agreement:** Since stakeholders often have strongly divergent political, social, and economic interests, the search for an equilibrium pushes ethicists to formulate the principles in very general and rather vague terms. Terms like ‘accountability’ or ‘fairness’ tend to be “thick in status, thin in content”,³³ functioning as mere “applause lights”³⁴ that signify tenuous political compromises rather than genuine ethical insights. We thus risk ending up with conceptually indeterminate, ‘toothless’ principles that do not specify what would count as a failure of fulfilment, and which are compatible with mutually exclusive technical practices.³⁵

3) **Regulatory capture and ethics washing:** Under the guise of democratically inclusive deliberation and social involvement, powerful political and economic actors may capture the process of standard-setting.³⁶ They may strategically pre-empt, weaken, or delay enforceable legislative regulations. At the same time, they may utilise the vaguely formulated principles by embracing them as a form of ethics washing or reputational management.³⁷

In sum, while the method of stakeholder-informed reflective equilibrium successfully generates a widely accepted and rationally justified vocabulary for AI certification, it does so at a cost: it risks obscuring deep normative conflicts, preserving the socio-economic status

quo, and leaving the most vulnerable stakeholders dependent on the self-regulatory goodwill of the very corporations whose power they would seek to constrain.

That being said, there are response strategies available to the proponents of the principles paradigm. In response to the first worry, they may emphasise that a properly conducted reflective equilibrium does not uncritically accept the pre-existing beliefs of stakeholders. Rather, it subjects them to continuous, iterative scrutiny by testing them against competing theories, empirical evidence, and counter-examples, thereby preventing the methodology from collapsing into mere conservative self-justification.³⁸ In response to the second and the third worry, proponents may point out that the ethical principles were never intended as either ‘done and dusted’ standards of concrete AI applications or as a substitute for concrete legislation and regulation. Rather, they are meant to serve as a shared normative vocabulary that can subsequently be specified into concrete, auditable criteria through standardised translational frameworks and be anchored in legislation with enforceable compliance measures.³⁹

2.3 Translation Assumption

2.3.1 The Principle-Practice Gap

The second basic assumption of the principles paradigm is that, once formulated, the general principles can be successfully translated into technical criteria for determining the ethical goodness or badness of AI applications in practice. In contrast to the task of principle formulation, the task of translation is explicitly recognised in the literature as a serious challenge, known as the *principle-practice gap*.⁴⁰ The gap refers to the disparity between, on the one hand, the *general ethical principles* formulated in abstract, humanistic, and qualitative terms in requirement catalogues, and on the other hand, the *practical*

³² See Cath 2016; Scanlon 2003

³³ Lee 2010; Bolte et al. 2022

³⁴ Max 2026, p. 2

³⁵ Max 2026; Mittelstadt 2019; Munn 2023

³⁶ Mensch et al. 2026; Pasetti 2026

³⁷ Dauvergne 2020; Redin and Ames 2026; Stilgoe 2019

³⁸ Cath 2016; Scanlon 2003

³⁹ Maclure and Morin-Martel 2024

⁴⁰ Corrêa et al. 2025b; Taddeo and Floridi 2018; Zhou and Chen 2023. See also the contributions in Hoffman and Bansal 2025

implementation of these principles in machine learning models that are trained on quantitative data sets using statistical objectives and mathematical parameters. Bridging this gap is the challenge of translating the general ethical principles into technical criteria for successful practical implementation.⁴¹

Although a number of translational frameworks have been proposed in recent years, there is a disconcerting tendency for the difficult task of actually translating principles into good practices to be “kicked down the road like the proverbial can”.⁴² The translation task is often left to the software developers, who are given little to no guidance on how to actually implement the general principles. What makes this tendency so disconcerting is the fact that, unlike other areas of ethical concern such as medical practice, AI development lacks a professional history, shared norms, fiduciary duties towards the public, and robust accountability mechanisms. In other words, the field of AI development lacks the professional background or culture that, in other fields, helps practitioners in responsibly translating general ethical principles into practice.⁴³ As a consequence, developers may translate and implement the same principles, not only in incompatible ways, but also in ways that prioritise efficiency and profit at the expense of the values actually intended by the requirement catalogues.⁴⁴

2.3.2 Frameworks of Principle-To-Practice Translation

The principles paradigm of AI certification thus requires a robust framework for translating general principles into technical criteria for assessing AI applications in practice. Without such a framework, the translation assumption is at best tenuous. Responding to this need, researchers, corporations, and NGOs alike have set out to develop and propose translational frameworks. Indeed, proposing such frameworks has become a rather active field of academic research, standard-setting, and corporate strategy, with a growing number of frameworks now available. These include:

Translational Ethics, proposed by Taddeo and Floridi (2018)

People + AI Guidebook, proposed by Google's multidisciplinary People + AI Research team (2019)

Ethics Framework, proposed by Digital Catapult (2021)

Ethics as a Service (EaaS), proposed by Morley et al. (2021)

Tools for Trustworthy AI, proposed by OECD (2021)

Trustworthy Artificial Intelligence Implementation (TAII), proposed by Baker-Brunnbauer (2021)

Responsible AI Standard, proposed by Microsoft (2022)

VCIO Framework (standing for Values, Criteria, Indicators, and Observables), proposed by the AI Ethics Impact Group (2020) and Peylo et al. (2022)

Ethical Problem Solving (EPS), proposed by Corrêa et al. (2025b)

Tripartite Ethical Policy Framework, proposed by Nakano (2026)

These frameworks vary considerably with respect to both the general principles to be translated and the specific means proposed to translate the former in practice. Yet they all share a common understanding of the principle-practice gap as, at bottom, an *engineering problem*: translating general principles into technical criteria for practical assessment, and thereby ensuring that AI applications are ethical, is simply a technical challenge that requires the right ‘toolbox’.⁴⁵ The shared core objective of these otherwise divergent frameworks is to present practitioners with such a toolbox to replace unreliable ad hoc translations with standardised and trustworthy translation procedures. Though the specific tools proposed by each framework vary, the toolboxes typically include:

⁴¹ Corrêa and Mönig 2024; Corrêa et al. 2025a. See also Barøe 2014

⁴² Mittelstadt 2019, p. 3

⁴³ Mittelstadt 2019; Munn 2023

⁴⁴ Corrêa et al. 2025a; Griffin et al. 2025

⁴⁵ Corrêa and Mönig 2024; Corrêa et al. 2025b; OECD 2021. See also Corrêa 2026; Ramachandran 2026

a) **Foresight and auditing tools:** Worksheets, impact assessment surveys, and qualitative troubleshooting questionnaires designed to map out risks prior to and during development.

b) **Procedural gates and checklists:** Iterative design-thinking exercises, stakeholder engagement matrices, and model documentation templates (such as model cards or data sheets) to ensure that compliance is actively recorded across product life cycles.

c) **Technical diagnostics:** Code repositories, bias-detection toolkits, and software-level intervention methods aimed at validating and correcting mathematical or algorithmic anomalies (such as model drift, opacity, or disparate impact).

By presenting practitioners with such standardised toolboxes, the translational frameworks are meant to guard the work towards creating ethical AI applications from unreliable and skewed translations of the ethical principles, and to turn this work into a standardised and reliable engineering practice.

2.3.3 Four Worries

For the principled approach to AI certification, the most pressing task is the development of a robust framework for translating general ethical principles into technical criteria for successful practical implementation. However, this effort faces at least four serious worries:

1) **Framework shopping:** The proliferation of divergent translational frameworks mirrors and intensifies the familiar problem of divergent requirement catalogues. Just as the early waves of AI ethics produced a ‘market for principles’,⁴⁶ where stakeholders could ‘shop’ for the most politically or reputationally convenient ones, the contemporary landscape runs the risk of degenerating into a similar ‘market for translations’. The unregulated development of differing translation frameworks allows political and economic actors to selectively adopt whichever one is least disruptive to their interests, thereby

enabling a sophisticated form of ethics washing under the guise of rigorous assessment.

2) **Underdetermination of successful translation:** It is underdetermined what actually counts as a successful translation.⁴⁷ Ethical principles are “essentially contested concepts” without a single, universally agreed-upon definition.⁴⁸ Consequently, it is possible to provide multiple, internally coherent, but mutually incompatible translations of the same principle. For example, in translating ‘fairness’, one developer might translate it into ‘demographic parity’ (which prioritises equal group outcomes), another into ‘equalised odds’ (which prioritises equal predictive accuracy). Both are internally coherent translations of fairness, yet they are mathematically incompatible.⁴⁹

3) **Normative displacement:** A closely related worry is that the translation frameworks shift the site of judgement about ethically good or bad AI applications into normatively loaded micro-decisions about measurement thresholds, baseline data selection, and acceptable risk parameters.⁵⁰ To enable practitioners to perform standardised translations, the frameworks must presuppose a host of normative background assumptions about what qualitative, ethical terms actually amount to in quantitative, technical ones. These background assumptions, in turn, become hardcoded into the AI applications by the practitioners, shielded from public contestation in virtue of being implicit in the apparently neutral tools of the framework.⁵¹

4) **Limited toolbox mentality:** A more general worry is that the underlying conception of ethical issues of AI applications as mere engineering problems fails to appreciate the complex and systemic character of these issues.⁵² Utilising statistical fairness metrics to adjust algorithmic weights does nothing to dismantle the structural and historical roots of social inequality that are imprinted in the data.⁵³ Similarly, calculating a model’s carbon footprint via tracking softwares like CodeCarbon is superficial, since it fails to account for the human and ecological costs of raw material extraction, premature

⁴⁶ Floridi 2019

⁴⁷ Max 2026; Mittelstadt 2019

⁴⁸ Mittelstadt 2019, p. 3

⁴⁹ Genovesi et al. 2024

⁵⁰ Max 2026; Mittelstadt 2019

⁵¹ Crawford 2021; Kong and Chang 2026; Mittelstadt 2019

⁵² Bolte et al. 2022; Bolte and van Wylsberghe 2024;

⁵³ Crawford 2021

hardware obsolescence, and the global distribution of electronic waste.⁵⁴ By framing complex social, economic, and ecological problems as discrete technical ones, the ‘toolbox mentality’ neglects the systemic character of tech-driven harm.

In sum, the translational frameworks proposed to make good on the translation assumption are far from uncontroversial. Though the frameworks may successfully provide reliable and technically feasible guidelines for bridging the principle-practice gap, they do so based on a host of normative background assumptions that surreptitiously shape the criteria of ethical AI applications, as well as the reduction of complex social, economic, and ecological problems to discrete technical ones. By reducing normativity to optimisation, the translational frameworks risk serving to disguise contested ethical-political issues as indisputable technical necessities.⁵⁵

Once again, there are response strategies available to the proponents of the principles paradigm. In response to the first worry, they may emphasise that ‘framework shopping’ is being pre-empted, at least partially, through international standards such as ISO/IEC 42001 and regulatory frameworks like the EU’s *AI Act*.⁵⁶ In response to the second worry, proponents may point out that what counts as a successful translation is context-dependent, meaning that the normative choice among mathematically incompatible translations of ethical principles will become clearly settled in the context of a specific domain, such as healthcare, law, or social media.⁵⁷ In response to the third worry, proponents may simply deny that standardised translational frameworks covertly hide normative background assumptions behind apparently neutral technical tools. They may point out that the frameworks explicitly mandate multi-stakeholder deliberation and ethics boards to ensure that baseline data selection, measurement thresholds, and risk parameters remain transparent, auditable, and subject to continuous expert and public contestation.⁵⁸ Finally, in response to the fourth worry, proponents may simply deny that the ‘toolbox

mentality’ is limited or problematic. They may instead emphasise this mentality as a virtue, one that makes the approach open to incorporating tools such as Life-Cycle Assessment (LCA) (see sec. 3.4) into the certification scheme in order to take the broader environmental, social, and economic impact of AI applications into account.⁵⁹

2.4 Arbitration Assumption

2.4.1 Value Monism and Value Pluralism

The third basic assumption of the principles paradigm is that conflicts arising between ethical principles can be rationally arbitrated. In other words, when ethical values conflict in specific AI applications, there is a reliable way of resolving the conflicts by determining which principles should restrict or override others. To properly understand this assumption, we must briefly take a step back and see it against the background of the meta-ethical discussion on value monism and value pluralism.⁶⁰

Value monism is the view that there is a single, ultimate ethical value or principle—such as ‘utility’ in utilitarianism or the ‘categorical imperative’ in Kantian deontology—to which all other ethical values can be reduced. On a monistic view, ethical arbitration is *theoretically straightforward*: when values appear to conflict, they can be reduced to the master value, thereby allowing a single correct decision procedure to resolve the conflict.

Value pluralism is the view that there are multiple, often incommensurable values—such as ‘autonomy’, ‘justice’, and ‘well-being’—that cannot be reduced to any other value. According to pluralism, conflicts between values are *not so easily resolved*, since satisfying one value often requires sacrificing or compromising another.

⁵⁴ Van Wynsberghe 2021

⁵⁵ See Feenberg 1999; Feenberg 2017

⁵⁶ EU 2024; ISO 2023

⁵⁷ Corrêa and Mönig 2026a; Brogle et al. 2025. See also sec. 2.4.2

⁵⁸ See Corrêa et al. 2025b

⁵⁹ See Corrêa et al. 2025b; Falk et al. 2026

⁶⁰ See Blum 2023; Heathwood 2015

The historical trajectory of AI ethics is that of a field which used to be dominated by monistic theories, but which underwent a structural shift towards pluralism.⁶¹ In the early decades of the field, most theories were value monistic, including notably Asimov's fictional 'three laws of robotics' (1950), Wiener's consequence-sensitive welfare framework (1988), and Maner's utilitarian computer ethics (1999). Over time, as AI applications transitioned from speculative fiction to actual technologies, the field underwent a decisive turn towards pluralism. Key milestones in this trajectory include: Beauchamp and Childress' pluralistic principlism (1979/2019), which profoundly shaped modern AI ethics; Moor's just consequentialism (1995; 2006), which combines deontological constraints of justice and rights with consequentialist calculations; and Dubljević and Racine's Agent-Deed-Consequence (ADC) model (2014), which integrates virtue ethics, deontology, and consequentialism to evaluate character, rules, and outcomes.

Today, value pluralism is undisputedly the dominant view in AI ethics and certification. Virtually every major AI governance framework presents a catalogue of multiple irreducible ethical principles.⁶² However, by subscribing to value pluralism, these frameworks all recognise, at least implicitly, that the principles they put forth can, and routinely do, come into conflict with one another.⁶³ Which raises the crucial question: how do we arbitrate in cases where principles conflict with each other? Consider for instance conversational AI companions designed to alleviate loneliness, where optimising for beneficence through simulated affective reciprocity risks creating a form of structural deception that directly compromises the user's autonomy and dignity. Similarly, consider a clinical AI recommendation system that may suggest an aggressive, high-risk drug dosage to maximise therapeutic beneficence, thereby coming into direct conflict with strict non-maleficence parameters and the patient's right to autonomous informed consent. To pretend that such

conflicts will resolve themselves, or that all the principles can somehow be equally respected afterall, is to invite the illusion of a "moral free lunch", namely, that all of the ethical principles can be simultaneously maximised without any costs or trade-offs.⁶⁴ If it is not to invite this illusion, or to leave the task of conflict resolution to the unguided ad hoc discretion of downstream software developers, a pluralistic principled approach to AI certification is obligated to provide a reliable method of arbitration in case of conflicts between principles.

2.4.2 Proposals for Principle Conflict Arbitration

However, when faced with the issue of conflicts between principles, proponents of the principles paradigm tend to resort to a series of vague proposals that fail to provide a reliable method of arbitration:

a) **The balancing metaphor:** A common response in the literature is that conflicting principles must be 'balanced' against one another.⁶⁵ This response, directly inherited from bioethical principlism, treats principles as *prima facie* duties that are ethically binding unless they conflict with an equal or weightier duty.⁶⁶ Yet, as Clouser and Gert famously argue in their critique of principlism, simply listing principles and instructing auditors to 'balance' them offers no systematic method for doing so.⁶⁷ In practice, the balancing metaphor functions as a rhetorical blank check: we are offered a checklist of values, but no objective weighing procedure. Downstream developers and auditors are left to rely on their own potentially inconsistent ad hoc judgements when deciding which principle must yield under which circumstances.⁶⁸

b) **Hierarchies and constraints:** Another response is to establish rigid lexical hierarchies or constraints, where certain principles, such as 'autonomy' or 'safety', are treated as inviolable, and thus cannot be overridden by calculations of utility or 'beneficence'.⁶⁹ This proposal is exemplified by Moor's "just consequentialism" and Gros et al.'s

⁶¹ The following overview of the historical trajectory of AI ethics is based on Harris and Dubljevic 2025

⁶² See EU 2024; IEEE 2019; OECD 2019/2024; UNESCO 2021; WHO 2021

⁶³ Corrêa and Mönig 2024; Corrêa and Mönig 2026a; Cremers et al. 2019. See also D'Acquisto 2020

⁶⁴ Max 2026, p. 2

⁶⁵ See EU 2024; Floridi and Cows 2019; Floridi et al. 2021; UNESCO 2021; WHO 2021

⁶⁶ Beauchamp and Childress 2019

⁶⁷ Clouser and Gert 1990

⁶⁸ Mittelstedt 2019. See also Corrêa et al. 2025a; Griffin et al. 2025

⁶⁹ See for instance IEEE 2019

“augmented utilitarianism”, where a set of deontological constraints function as initial screening conditions.⁷⁰ However, establishing a universally accepted hierarchy of principles has proven to be not only highly contentious, but in practice politically impossible.⁷¹ A hierarchy of principles or context-insensitive constraints risks introducing a level of rigidity into the framework that prevents it from accommodating situations where, for example, a minor infraction of one principle could prevent a major loss of aggregate welfare.

c) Methodological contextualism: The most defensible view found in the academic literature and among the available AI governance frameworks seems to be a form of methodological contextualism. This view assumes that the relative weights of the different principles, and the specific trade-offs allowed in any given scenario, are contextually determined by the facts of a specific AI application.⁷² A good example of this method in action is Mönig and Corrêa's *Catalogue of Field-Specific Ethical Requirements for AI Certification*, which systematically outlines how overarching principles should be weighted and which trade-offs are permissible in specific fields like arts, law enforcement, and social media.⁷³ The strength of such a contextual approach lies in its flexibility and sensitivity to domain-specific risks and duties. However, a contextualist approach also runs the risk of simply delegating the hard task of actual arbitration downstream. By asserting that weights and trade-offs are contextually determined, principled approaches avoid making difficult normative commitments at the framework level, once again effectively ‘kicking the can down the road’.⁷⁴

In sum, the varied proposals for principle conflict arbitration reveal that the principles paradigm suffers from a persistent problem of “underdetermination with respect to conflict”.⁷⁵ General ethical principles do not come with clear rules of priority, and the catalogues of these principles typically do not offer such priority rules or general ways of arbitrating conflicts among the principles. The closest thing

to a reliable method of arbitration that the principled approach has to offer seems to be a localised, contextualised application of the same method of stakeholder-informed reflective equilibrium that is used to formulate the ethical principles in the first place. In practice, this means that the task of resolving difficult ethical and political trade-offs is left to consensus-seeking deliberations in localised development settings. Because these local compromises are unconstrained by clear, systematic arbitration procedures, they remain highly underdetermined and, as such, prone to capture by political and economic actors that seek to further their own strategic and commercial interests.

2.5 Isolation Assumption

2.5.1 Individual Artefact as Unit of Assessment

The fourth basic assumption of the principles paradigm is that an AI application can be ethically assessed in relative isolation from the broader ecological, social, and economic system as part of which it is developed and deployed. On the principled approach, the *basic unit of ethical assessment* is the AI application viewed as an individual artefact. The AI application is approached as a discrete, self-contained, and static entity whose ethically relevant characteristics can be identified relatively independently of the broader ecological, social, and economic contexts of its development and use.⁷⁶ In other words, the ethical profile of the technology is taken to reside within the boundaries of the individual model or software application itself, and thus to have nothing to do with the material and social infrastructures upon which it relies.

Consequently, ensuring the ethical goodness of AI applications is viewed primarily as an *engineering problem*: a technical challenge of reining in the ‘sharp edges’ of a given model by optimising its technical parameters, adjusting algorithmic weights, and securing safe data inputs.⁷⁷ All of

⁷⁰ Gros et al. 2025; Harris and Dubljevic 2025; Moor 1995; Moor 2006

⁷¹ Max 2026; Mittelstadt 2019. See also Maclure and Morin-Martel 2024

⁷² See EU 2024; OECD 2021; UNESCO 2021; WHO 2021

⁷³ Corrêa and Mönig 2026a. See also Brogle et al. 2025

⁷⁴ Mittelstadt 2019

⁷⁵ Max 2026, p. 3. See also Mittelstadt 2019

⁷⁶ Bolte et al. 2022; Bolte and van Wynsberghe 2025; Hagendorf 2020; Sætra 2021

⁷⁷ Corrêa et al. 2025b; Corrêa 2026; Ramachandran 2026

these interventions, so the idea goes, can be successfully performed on individual AI applications in the relative vacuum of programmer workshops and development laboratories. By extension, ethical auditing and certification procedures are designed to evaluate compliance on a case-by-case or application-by-application basis. If an individual AI application can be validated for domain-specific technical criteria for accountability, fairness, robustness, and so on, it is deemed ethically compliant and awarded a stamp of approval.⁷⁸

2.5.2 Two Worries

The assumption that an individual AI application can be ethically assessed and certified in relative isolation from its broader social, economic, and ecological context is beset by serious worries. Two worries in particular deserve to be emphasised:

1) Blind to cumulative macro-impacts: By focusing exclusively on the design and performance of isolated AI applications, the principled approach is blind to cumulative macro-impacts.⁷⁹ An individual machine learning model can be optimised to be highly efficient in a local setting, yet when such models are deployed at scale across society, they trigger macro-level changes that may counteract local benefits. An example of this problem is the so-called *rebound effect*, where incremental gains in algorithmic or hardware efficiency lower the cost of computation and expand its accessibility, which in turn drives up the aggregate demand for data centres, model training, and continuous scaling, thereby accelerating the sector's overall environmental footprint.⁸⁰ Because the principled approach restricts its focus to individual, isolated artefacts, it neglects such macro-level dynamics. It evaluates the 'trees' of immediate software performance, while failing to see the 'forest' of an expanding, resource-intensive planetary infrastructure.⁸¹

2) Concealing environmental and human costs: By viewing the individual AI application as the basic unit of assessment, the principled approach tends to conceal the

human and environmental costs of these technologies. Despite what metaphors like 'the cloud' and 'virtual assistants' might suggest, AI is not an ethereal, disembodied phenomenon.⁸² Rather, it depends on physical data centres, mineral extraction, global supply chains, toxic waste disposal, and cheap labour, all on a massive scale. For example, the production of specialised GPUs and batteries requires the extraction of rare-earth elements, lithium, and cobalt, which takes place under conditions that often involve slavery-like labour, soil pollution, and severe water depletion in local communities.⁸³ Another example is how the development of AI systems relies on a massive, highly precarious, and underpaid workforce of data annotators.⁸⁴ By taking the individual AI application as its basic unit of assessment, the principled approach is able to treat such facts as 'externalities' that are divorced from the ethical profile of the application itself. Consequently, an AI application that relies on ethically problematic infrastructures may still become certified as a 'trustworthy AI', as long as it conforms to the narrow technical criteria spelled out in a given certification scheme.

In sum, the isolation assumption is a deeply problematic assumption. By viewing individual AI applications in abstraction from their social, economic, and ecological contexts, principles-based certification schemes risk concealing systemic harms and thereby helping to insulate corporate power from critique. Assessing AI applications in isolation allows developers and standard-setting bodies to award stamps of ethical compliance to applications that are, in virtue of the material and social infrastructures on which they rely, deeply unjust and unsustainable.

However, there are ways for the proponents of the principles paradigm to respond to these worries and defend the isolation assumption. In response to the first worry, proponents may argue that certification is fundamentally a product-level conformity mechanism designed to hold specific developers and deployers accountable for issues within their direct operational control. Macro-level dynamics reflect broad socio-

⁷⁸ See Corrêa and Mönig 2026a; Genovesi and Mönig 2022; Genovesi et al. 2024

⁷⁹ Bolte et al. 2022; Bolte and van Wynsberghe 2025; Sætra 2021

⁸⁰ Chouksey et al. 2026; Dauvergne 2020

⁸¹ Robbins and van Wynsberghe 2022

⁸² Crawford 2021

⁸³ Dauvergne 2020; Falk et al. 2026; Fuchs 2014

⁸⁴ Mensah and van Wynsberghe 2025; Waelen and van Wynsberghe 2025

Philosophical Foundations of AI Certification

economic trends that cannot be governed through individual product testing alone. Hence, rather than overburdening individual product audits, these macro-impacts are more effectively addressed through overarching legislative frameworks such as the EU's *Green Deal*. In response to the second worry, proponents may contend that taking the individual AI application as the unit of assessment does not require treating supply-chain harms, mineral extraction, or precarious data labour as

mere 'externalities'. In fact, contemporary AI certification schemes increasingly integrate life-cycle metrics, supply-chain tracking, and social labour audits into their compliance requirements.⁸⁵ In this way, AI certification can enforce transparency about, and accountability for, systemic issues without sacrificing the advantages of clearly defined audit boundaries.

⁸⁵ Corrêa and Mönig 2024; EU 2024; ISO 2023

3 An Alternative Paradigm: Systems

3.1 The General Idea of Systemic AI Certification

In recent years, the dominant principles paradigm in AI ethics has increasingly been challenged by what may be called the systems paradigm. This paradigm is being promoted particularly by researchers in Sustainable AI⁸⁶ and critical AI studies,⁸⁷ and it is inspired by recent developments in social philosophy,⁸⁸ philosophy of technology,⁸⁹ and complex systems theory.⁹⁰ It shifts the focus of assessment from AI applications viewed as isolated artefacts to the complex social, economic, and ecological systems in which they are embedded. And instead of applying abstract principles in a top-down manner, it evaluates the impact of AI applications on the functioning of these systems in light of concrete standards already operative in them. As such, the systems paradigm constitutes a direct response to the perceived shortcomings of the principles paradigm.

Despite the growing prominence of the systems paradigm in both foundational and interdisciplinary research, there are no full-fledged systemic certification schemes yet available. The vast majority of current certification frameworks remain firmly anchored in the principles paradigm. Nevertheless, given that the systems paradigm is a serious alternative to the dominant principled approach in the field of AI ethics, and given that AI certification is essentially an institutionalised version of AI ethics, it is worth exploring and articulating the basic assumptions of a systemic approach to AI certification.

The systems paradigm is based on the idea that AI applications must always be viewed in relation to the complex social, economic, and ecological systems in which they are embedded.⁹¹ These systems are understood, not as normatively neutral substrates, but rather as historically

emergent, normatively constituted contexts of processes and practices, governed by implicit norms of good functioning.⁹² Accordingly, the ethical assessment of an AI application focuses, not on its conformity to a predefined set of ethical principles, but rather on its impact on or contribution to the functioning of the complex system in which it is embedded. This assessment procedure begins with mapping the relevant system and explicating its implicit norms.⁹³ The auditors then diagnose the AI application's concrete impacts on the system, tracing how its life cycle interacts with, reproduces, modifies, or transforms the system, assessing not only direct technical inputs and outputs, but also indirect, cumulative, and long-term effects.⁹⁴ On this basis, an AI application may be certified only if it can be systemically justified, that is, if its socio-technical and material life cycle contributes positively to, or at least does not actively undermine, the well-functioning of the complex system of which it forms a part.

Like the principles paradigm, the systems paradigm relies on significant basic assumptions. Specifically, any future systemic AI certification scheme must operate with the following four foundational assumptions:

Holism assumption: an AI application is only ethically assessable relative to the complex social, economic, and ecological system in which it is structurally embedded.

Mapping assumption: the complex system within which an AI application is embedded can be mapped.

Diagnosis assumption: the impact of an AI application on the complex system can be reliably diagnosed.

Evaluation assumption: an AI application's systemic impact can be reliably evaluated in light of the complex system's implicit norms.

⁸⁶ Bolte and van Wynsberghe 2025; Falk and van Wynsberghe 2024; Khakurel et al. 2018; van Wynsberghe 2021; van Wynsberghe et al. 2023

⁸⁷ Coeckelbergh 2022; Coeckelbergh and Sætra 2023; Lindgren 2023a; Lindgren 2023b

⁸⁸ Celikates 2018; Feenberg 1999; Feenberg 2017; Fuchs 2014; Jaeggi 2018

⁸⁹ Bosschaert and Blok 2026; Lemmens et al. 2017; Stellinga et al. 2025; Verbeek 2021

⁹⁰ Hooker 2011; Jørgensen et al. 1992; Thurner et al. 2018; Tranquillo 2019

⁹¹ Bolte et al. 2022; Bolte and van Wynsberghe 2025; Crawford 2021; Onder 2026; Robbins and van Wynsberghe 2022; Sætra 2021

⁹² Celikates 2018; Jaeggi 2018. See also McShane 2004

⁹³ Crawford 2021; Crawford and Joler 2018

⁹⁴ Falk et al. 2026; Plociennik et al. 2025

Before turning to elaborate and problematise these basic assumptions, it is worth noting a further respect in which the systems paradigm would change the nature of AI certification. Under the principles paradigm, a certificate is typically awarded as a one-off stamp of approval, treating an AI application as a finished product that consumers can henceforth trust. However, because complex systems are inherently dynamic, a systemic approach to AI certification would regard such one-off audits as structurally inadequate. Instead, certification would have to take the form of a continuous, dynamic process involving periodic inspections throughout an AI application's entire life cycle. This model can be compared to the health and safety inspection schemes applied to restaurants. A restaurant is not certified as safe once and for all when its kitchen is built. Rather, it must pass regular, unannounced inspections by public health authorities, who audit its day-to-day practices, supply chains, and hygiene standards under real-world operating conditions. Similarly, systemic AI certification would require continuous monitoring of the impacts of an AI application on the complex social, economic, and environmental system of which it forms a part. If an inspection were to reveal that a previously certified AI application had generated unforeseen distortions or systemic harms, its certificate could be suspended, prompting repeated calls for technical modifications or broader structural changes to the system itself.

3.2 Holism Assumption

3.2.1 Complex System as Unit of Assessment

The first basic assumption of the systems paradigm is that an AI application is only ethically assessable relative to the complex social, economic, and ecological system in which it is structurally embedded. In contrast to the principles paradigm, the systems paradigm emphasises that the development and operation of AI applications inevitably rely on a vast, global infrastructure of material and social processes. On the material level, training and running

contemporary AI applications require large-scale extraction of metals and rare-earth minerals, energy-intensive data centres, substantial water withdrawals for thermal cooling, and the eventual disposal of toxic electronic waste.⁹⁵ On the social level, AI applications rely on global pipelines of precarious human labour, including the 'ghost labour' of crowdsourced data annotators, data scrapers, and content moderators.⁹⁶ All such structural dependencies, so the systemic idea, are just as integral to the ethical profile of an AI application as its algorithmic logic or statistical performance.

Accordingly, the systems paradigm takes the basic unit of ethical assessment to be, not the individual AI application, but rather the *complex system of ecological, social, and economic processes* in which it is embedded. These systems are understood as dynamic orders of human and non-human co-existence, encompassing the structured interactions of manifold and diverse elements, including organisms and natural environments, individuals and groups, social institutions and technologies. The complex systems are analysed into three interconnected spheres:⁹⁷

Ecological sphere: encompassing organisms, ecosystems, climate cycles, and raw material flows.

Social sphere: encompassing human communities, social practices and institutions, labour relations, cultural orientations, and structures of social cohesion.

Economic sphere: encompassing processes of commodity production, resource allocation, and market exchange.

That being said, in real-world cases, these three spheres are tightly interwoven and continuously interact. When an AI application is introduced as a new commodity in the economic sphere, it draws on raw materials and energy from the ecological sphere and affects labour practices and power asymmetries in the social sphere. Accordingly, to ethically assess an AI application requires us to evaluate the impacts of its entire life cycle across all three spheres, which is to say, on the complex system as a whole.

⁹⁵ Crawford 2021; Dauvergne 2020; Falk et al. 2026; Falk and van Wynsberghe 2024

⁹⁶ Crawford 2021; Graham et al. 2017; Mensah and van Wynsberghe 2025; Waelen and van Wynsberghe 2025

⁹⁷ Bolte et al. 2022; Bolte and van Wynsberghe 2025; Waelen and van Wynsberghe 2025

3.2.2 Implicit Normativity of Complex Systems

A key feature of the systems paradigm's conception of complex systems is that these are not neutral or value-free substrates that provide the resources needed to develop and deploy AI technologies. Rather, complex eco-socio-economic systems are *normatively constituted*.⁹⁸ The idea is this: we understand a given system—whether an education system, an ecosystem, or a labour market—not merely by listing its various elements, but by observing how the system functions as an organised whole. Specifically, we look at what *purpose(s)* the interactions of a system's elements serve. For example, the goings-on of an education system serve the purpose of cultivating human social and cognitive capacities; those of an ecosystem serve the purpose of preserving the biospheric carrying capacity that supports present and future life; and those of a labour market serve the purpose of securing the dignified livelihood of individuals and productive economic cooperation. Crucially, there is no neutral or value-free ways of achieving these purposes, only better or worse ones. That is to say, the goings-on of a system are governed by *implicit norms* of what counts as *good functioning*. To return to the examples, teacher-pupil interactions are governed by implicit norms of critical inquiry, pedagogical care, and open dialogue; ecosystem interactions are governed by norms of ecological integrity, biophysical resilience, and regenerative balance; and labour market interactions are governed by norms of fair compensation and non-exploitative cooperation. We can only properly understand the interactions of a system's elements, and by extension the system as a whole, with reference to such constitutive standards of good functioning or “ethical-functional norms”.⁹⁹

This recognition of implicit yet concrete norms already operative in the very systems we seek to assess marks a fundamental departure from the principles paradigm. Whereas the principles paradigm seeks to govern AI applications by imposing general ethical principles from the

outside, the systems paradigm recognises that context-specific norms of good functioning are *always already operative* within the host systems themselves. Although these norms can only be properly specified in relation to actual, concrete systems, we may roughly specify the kinds of norms that govern the three aforementioned spheres:

Ecological sphere: governed by norms of ecological integrity, biophysical resilience, and planetary boundaries.¹⁰⁰

Social sphere: governed by norms of social cohesion, human dignity, equity, and democratic self-determination.¹⁰¹

Economic sphere: governed by norms of sustainable resource allocation, fair distribution, and non-exploitative labour.¹⁰²

The crucial point is that these norms do not constitute an external checklist along the lines of the familiar catalogues of minimal requirements. Rather, they are internal to the very ways in which the different spheres function and in which their elements interact with each other.¹⁰³ They are the norms that make a complex system what it is, namely, a system with a specific (set of) purpose(s) that it purports to achieve well.

Accordingly, on a systemic approach, an AI application can only be ethically assessed as an integral part of a complex system. This is so, not only because the AI application is embedded in and depends upon the broader ecological, social, and economic infrastructure, but also because it is the complex system itself that supplies the very standards or norms in terms of which the AI application must be evaluated. An AI application cannot be ethically assessed on its own as an isolated artefact, because its ethical profil depends entirely on whether its socio-technical life cycle supports and enhances, or disrupts and degrades, the well-functioning and long-term viability of the overarching system of which it forms a part.

⁹⁸ The following reconstruction of the implicit normativity of complex systems is adopted from Jaeggi 2018, ch. 3. See also Celikates 2018; McShane 2004; Neuhauser 2018

⁹⁹ Jaeggi 2018, p. 113. See also Neuhauser 2018

¹⁰⁰ McShane 2004; Stellinga et al. 2025; van Wynsberghe 2021

¹⁰¹ Celikates 2018; Coeckelbergh and Sætra 2023; Feenberg 1999; Feenberg 2017; Mensah and van Wynsberghe 2025; Waelen 2022; Waelen and van Wynsberghe 2025

¹⁰² Crawford 2021; Halsband 2022; Orm 2026; Waelen and van Wynsberghe 2025

¹⁰³ Bolte et al. 2022; Jaeggi 2018; McShane 2004

3.2.3 Advantage and Two Worries

By shifting the focus from isolated artefacts to complex eco-socio-economic systems as the basic unit of assessment, the systems paradigm directly responds to and seeks to overcome the two worries raised against the principles paradigm's isolation assumption. By expanding the analytical horizon beyond individual artefacts to the complex systems in which they are embedded, the holism assumption enables a systemic approach, firstly, to capture the cumulative macro-impacts of the development and deployment of AI applications. Secondly, the holism assumption enables a systemic approach to make explicit the vast supply chains, and thereby the great environmental and social costs, involved in the development and deployment of AI applications.¹⁰⁴

Despite this advantage, taking complex systems as the basic unit of assessment invites two serious worries:

1) Incompatible with certification: The first worry is that complex eco-socio-economic systems are simply too vast, too dynamic, and thus too poorly bounded to be subjected to certification. Traditional certification regimes rely on clear boundaries, standardised testing environments, and static compliance thresholds applied to discrete products before market entry. If the unit of assessment encompasses planetary extraction networks, fluctuating energy grids, and evolving social practices, the audit boundary, so the worry goes, appears to expand infinitely, rendering certification practically impossible.

2) Ontologically spooky: Another, more theoretical worry is that complex systems are ontologically spooky or strange.¹⁰⁵ Talking about complex systems as the basic unit of assessment seems to suggest that they are a special class of normatively constituted macro entities existing over and above the individual elements of which they consist. Yet, unlike its elements, such macro entities lack clear empirical boundaries and are therefore difficult to study. To treat complex systems as real and, indeed, normatively constituted entities, so the worry goes, would be to impose

an artificial unity on what is merely an ad hoc collection of discrete physical objects and human choices.

We shall return to the first worry again below (see sec. 3.3.3). At present, it will suffice to note that this worry seems to rest on the false premise that certification must take the form of a one-time, static, pre-deployment stamp of approval, which is far from the case. Consider for example health and safety certification of restaurants or the environmental impact certification of companies.

As for the second worry, proponents of the systems paradigm can point out that the existence of macro entities, and normatively constituted entities, is widely accepted across both the natural and the social sciences. For example, in ecology and environmental science, ecosystems are routinely analysed as complex, self-organising systems that possess literal, non-metaphorical health structures and functional dependencies, even though their spatial boundaries are interdetermined and defined relative to research scales.¹⁰⁶ Similarly, economists and sociologists routinely study macroeconomic networks, labour markets, and urban socio-ecological systems that cannot be understood merely by aggregating micro-level individual transactions. These systems exhibit emergent systemic behaviours, feedback loops, and structural crises that operate 'behind the backs' of individual agents.¹⁰⁷ Ultimately, the justification for adopting the complex system as the basic unit of ethical assessment is that it yields greater explanatory power than restricting oneself to the individual artefact. It enables AI ethicists and auditors to explain and respond to real-world harms—such as rebound resource depletion, global e-waste accumulation, and structural labour precarity—that a strictly artefact-focused approach would remain blind to.

¹⁰⁴ Bolte et al. 2022; Bolte and van Wynsberghe 2025; Crawford 2021; Sætra 2021

¹⁰⁵ This worry is adopted from Mackie's so-called 'argument from queerness'. See the classic formulation in Mackie 1977. See also Brink 1984; Garner 1990; Olson 2014, ch. 5 and 6

¹⁰⁶ McShane 2004; Odenbaugh 2011

¹⁰⁷ Gomes and Gubareva 2020; Jaeggi 2018

3.3 Mapping Assumption

3.3.1 Systems Cartography 1: Elements, Relations, Dynamics, and Boundaries

The second basic assumption of the systems paradigm is that the complex system within which an AI application is embedded can be mapped, at least to a relevant degree of accuracy. To evaluate an AI application's impact on the functioning of a host system, auditors must first be able to construct a comprehensive 'map' of that system. This is the task of what may be called *systems cartography*: to gather and organise all relevant information about a given eco-socio-economic system into an integrated visual-conceptual representation.¹⁰⁸ Perhaps the best example of this is Crawford and Joler's "Anatomy of an AI System", which is a detailed representation of the complex system in which the Amazon Echo Dot is embedded.¹⁰⁹

In the first instance, systems cartography sets out to map the following four dimensions of a complex system:

a) Elements: The map must identify the system's elements, at different scales and across the aforementioned ecological, social, and economic spheres. At the *micro-level*, this includes organisms, human actors, and discrete technical components; at the *meso-level*, it includes ecosystems, resource deposits, and companies; and at the *macro-level*, it includes climate, labour markets, and institutional legal frameworks.¹¹⁰

b) Relations: The map must trace the diverse relations between the system's elements, both within and across the three spheres. These include *causal interactions*, such as data centre energy demands causing an increase in regional grid emissions; *functional dependencies*, for example the way an algorithmic model's operational accuracy depends on human content moderation or specialised thermal cooling systems; and *emergent properties*, such as systemic power asymmetries, environmental lock-ins, or institutional bias, which cannot be reduced to relations among singular elements but arise from the interactions of the whole.¹¹¹

c) Dynamics: The map must also chart feedback loops and other non-linear dynamics that unfold over time. Key examples include the aforementioned *rebound effect*, where gains in computational or hardware efficiency lower processing costs, inadvertently driving up the aggregate demand for data centres, model training, and energy; and *user-system feedback loops*, where adaptive machine learning models alter user behaviour, which in turn generates new training data that reshapes subsequent model outputs.¹¹²

d) Boundaries: Finally, the map must also draw the spatial, temporal, and organisational boundaries of the system. In *ecological terms*, this involves setting so-called Life-Cycle Assessment (LCA) inventory boundaries, such as Scope 1, 2, and 3 emissions spanning raw material extraction, hardware manufacturing, distribution, use, and e-waste disposal (we shall return to this in sec. 3.4); in *social and economic terms*, it involves drawing jurisdictional and sectoral boundaries.¹¹³

Notably, to construct such a systems map is an *interdisciplinary endeavor*.¹¹⁴ It requires insights that can only be gained by combining methods from multiple disciplines, including computer science, environmental engineering, supply-chain logistics, critical political economy, science and technology studies (STS), and philosophy. Only through such interdisciplinary collaboration can auditors make visible the full socio-material realities of AI applications.

3.3.2 Systems Cartography 2: Explicating Implicit Norms

Since the complex systems in which AI applications are embedded are normatively constituted, the second and indispensable aspect of systems cartography is the explication of a system's implicit norms of good functioning. This is an *interpretative endeavour*, yet not one that is detached from or secondary to the mapping of a system's dynamically interrelated and interacting elements. For in doing so, we are always already interpreting these various

¹⁰⁸ Bamgoye and Avellán 2025; Castellani et al. 2025; Crawford 2021

¹⁰⁹ Crawford and Joler 2018

¹¹⁰ Crawford and Joler 2018; Wang and Blok 2026

¹¹¹ Chouksey et al. 2026; Robbins and van Wynsberghe 2022; Waelen 2022; Waelen and van Wynsberghe 2025

¹¹² Chouksey et al. 2026; Coeckelbergh and Sætra 2025; Mills and Sætra 2025; Sætra 2021

¹¹³ Bolte et al. 2022; Falk et al. 2026; Onder 2026; Raworth 2012; Röckström et al. 2009

¹¹⁴ See the contributions in van Wynsberghe et al. 2023

elements as forming norm-governed processes or subsystems that serve specific purposes. As such, the task of systems cartography would be left incomplete without the explication of the system's implicit norms.

The general idea is simple enough: we first explicate the *purposes* that specific processes or subsystems serve, and then explicate the *implicit norms* that govern these processes or subsystems from within, defining what counts as good versus bad functioning.¹¹⁵ We can illustrate this procedure with two examples:

Processes of power draw, power distribution, and server cooling on which data centres depend

What we must explicate here is what allows us to understand these processes as contributing to the functioning of a data centre. This involves, firstly, recognising that they serve the *purpose* of ensuring effective and stable data processing; and secondly, identifying the *implicit norms* that govern them: energy efficiency, biophysical resource conservation, and local hydrological balance. These are the norms that define what counts as a well-functioning data centre.

Processes of rare-earth mineral mining, chemical refining, and semiconductor fabrication on which the hardware manufacturing industry relies

Again, we must explicate what allows us to understand these processes as contributing to the functioning of the hardware manufacturing industry. This involves, firstly, recognising that they serve the *purpose* of supplying the specialised hardware required to run AI models; and secondly, identifying the *implicit norms* that govern them: worker safety, non-exploitative labour, environmental remediation, and supply-chain justice. These are the norms that define what counts as a well-functioning hardware manufacturing industry.

Explicating these implicit norms amounts to explicating the *conditions of fulfilment* of the system and its constituent processes and subsystems, that is, the conditions of successful reproduction of the system over time.¹¹⁶

Successful reproduction means that the system maintains its internal structure over time and continues to fulfill its current purposes. But it also means that the system keeps multiple pathways of evolution open, allowing it to adapt and transform in response to shifting social needs and ecological feedback, and not become trapped in path-dependent environmental or institutional lock-ins.¹¹⁷

Hand in hand with this explication of fulfilment conditions goes the explication of the *minimal conditions of system reproduction*. These represent the non-negotiable ecological, social, and economic preconditions without which the complex system could not continue to function as a system, that is, to reproduce its internal structure over time. In the ecological sphere, these minimal conditions are set by biophysical carrying capacities and planetary boundaries, such as greenhouse gas concentration limits, freshwater withdrawal caps, and chemical pollution thresholds.¹¹⁸ In the social and economic spheres, they encompass basic human rights, non-exploitative labour conditions, community health, and equitable resource distribution.¹¹⁹ Explicating these minimal conditions identifies the critical tipping points and limits beyond which we would face systemic collapse.

3.3.3 Advantage and Two Worries

With the procedure of mapping of complex systems, and specifically of explicating their implicit norms of good functioning, the systems paradigm responds to and seeks to overcome the shortcomings of the principles paradigm's formulation and translation assumptions. The mapping assumption replaces the top-down formulation and checklist translation of general ethical principles with the interdisciplinary endeavour of mapping the concrete goings-on of a system and to explicate the ethical-functional norms that govern them. As such, the systems paradigm avoids the worries about the underdetermination of principles and translations, normative displacement, and a limited toolbox mentality.

Despite these advantages, the mapping assumption is a weighty one, and it gives rise to two serious worries:

¹¹⁵ Celikates 2018; Jaeggi 2018; Neuhauser 2018

¹¹⁶ Jaeggi 2018, p. 113

¹¹⁷ Bolte et al. 2022; Bolte and van Wynsberghe 2025; Robbins and van Wynsberghe 2022

¹¹⁸ Bolte et al. 2022; Brevini 2023; Falk et al. 2026; Onder 2026; Stellinga et al. 2025

¹¹⁹ Crawford 2021; Mensah and van Wynsberghe 2025; Waelen 2022; Waelen and van Wynsberghe 2025

1) System unboundedness: The first worry is that it is practically impossible to draw the boundaries of a complex system in a non-arbitrary manner. Such systems do not possess self-evident boundaries or natural cut-off points. Indeed, when we set out to construct a comprehensive map of a system in which an AI application is embedded, we seem to face two infinite regresses. When we try to map the material and human supply chains required to build, train, and run an AI application, we seem to face an *upstream regress* of virtually endless supply-chain dependencies. Similarly, when we try to map the effects of deploying an AI application, we seem to face a *downstream regress* of an equally endless web of socio-political interactions and environmental impacts unfolding over time. This creates a dilemma for systemic AI certification. If auditors refuse to set arbitrary boundaries, systems cartography becomes an impossible quest to map the entire global infrastructure. Conversely, if auditors do set arbitrary boundaries to make the audit manageable, they risk excluding upstream supply-chain dependencies and downstream effects, thereby reintroducing the worries associated with the isolation assumption of the principles paradigm (see sec. 2.5.2).¹²⁰

2) Projection of norms: The second worry is that we might project our own subjective norms and values onto the complex systems and their constituent processes and subsystems.¹²¹ That is, how can we be sure that auditors are genuinely explicating a system's implicit norms, rather than projecting their own subjective values, cultural expectations, or political biases onto it?¹²² For example, an auditor with neoliberal convictions might view an agricultural system's implicit norms as crop yield optimisation and financial gain, whereas an environmentalist auditor might instead view them as biophysical resilience and planetary health. If the explication of a system's conditions of fulfilment is merely a covert vehicle for personal or institutional normative preferences, then the systems paradigm fails to provide a reliable basis for the ethical assessment of AI applications.

Both of these worries highlight critical pitfalls of systems cartography. Nevertheless, proponents of the systems paradigm can respond to them by proposing a set of methodological safeguards. The worry of system unboundedness may be assuaged by pointing out that system boundaries need not be self-evident or natural in order to be non-arbitrary and objective. There are several non-arbitrary ways of drawing such boundaries available, including adopting LCA inventory boundaries that establish explicit, standardised evaluative cut-offs,¹²³ or anchoring the mapping in non-negotiable biophysical and social ceilings, such as planetary boundaries and the Doughnut Economics framework.¹²⁴

The worry of normative projection, in turn, may be assuaged by emphasising that the explication of implicit norms is not a top-down exercise conducted by isolated experts. Furthermore, the risk of subjective projection may be mitigated through participatory infrastructures that enable affected communities and domain experts to contribute to the process of defining, verifying, and contesting the system's implicit norms. It may also be mitigated through independent oversight or meta-auditing bodies to “audit the auditors themselves”,¹²⁵ ensuring that the process of norm explication remains transparent, subject to peer-review, and protected against regulatory capture and individual bias.

3.4 Diagnosis Assumption

3.4.1 Life-Cycle Assessment I: Overarching Framework

The third basic assumption of the systems paradigm is that the impact of an AI application on the complex system can be reliably diagnosed. In other words, there is a robust methodology or framework for diagnosing an AI application's systemic impact. The systems paradigm assumes that this framework is *Life-Cycle Assessment (LCA)*.

¹²⁰ See Crawford 2021; Crawford and Joler 2018; Polk et al. 2026

¹²¹ This worry is adopted from Hume's, and later Blackburn's, account of how the human mind tends to project subjective impressions onto objective reality. See Hume 1998; Hume 2011; Blackburn 1984, ch. 6; Blackburn 1998. See also Boehm 2021

¹²² Pasetti 2026

¹²³ See Falk et al. 2026; ISO 2006a; ISO 2006b; ITU 2024

¹²⁴ See Bolte et al. 2022; Falk et al. 2026; Onder 2026; Raworth 2012; Röckström et al. 2009

¹²⁵ Pasetti 2026, p. 23

Derived from environmental engineering and standardised by international protocols (such as ISO 14040, ISO 14044, and ITU-T L.1410), LCA is a comprehensive methodology for assessing the systemic impacts of a product or service across every stage of its life cycle. Although originally developed for environmental impact assessments, LCA has come to be seen as the overarching framework for diagnosing the systemic impact of AI applications across the ecological, social, and economic spheres.¹²⁶

To perform a comprehensive cradle-to-grave impact assessment of an AI application, auditors must examine it across the five interconnected *stages of its life cycle*:¹²⁷

S1) Raw material extraction: mining of lithium, cobalt, copper and rare-earth minerals.

S2) Hardware manufacturing: semiconductor fabrication, GPU assembly, and server production.

S3) Logistics and distribution: global air and freight transport of specialised hardware.

S4) Use and operation: algorithmic model training, fine-tuning, and real-world deployment.

S5) End-of-life processing: decommissioning of hardware, informal dismantling, and e-waste collection and disposal.

To systematically diagnose an AI application's systemic impact across these five stages, the LCA framework classifies impacts along two formal axes. Along the first axis, it distinguishes between three *orders of impact*:

O1) First-order impacts: the direct biophysical and computational footprint of the hardware and infrastructure required to build, run, and dispose of an AI application. Examples: data centre electricity consumption, operational greenhouse gas emissions, water withdrawals for server cooling, and e-waste generation.¹²⁸

O2) Second-order impacts: the indirect operational, environmental, and social changes induced by deploying AI applications in specific real-world contexts. Examples:

industrial energy optimisation, precision agricultural yield forecasting, automated medical diagnostic triage, and supply-chain logistics streamlining.¹²⁹

O3) Third-order impacts: long-term systemic transformations, behavioural feedback loops, and structural realignments that emerge over time as a result of widespread adoption of AI applications. Examples: rebound effects, consumerist speed-up, market power centralisation, and path-dependent infrastructural lock-ins.¹³⁰

Along the second axis, the LCA framework diagnoses how these impacts manifest across three different *levels of analysis*:¹³¹

L1) Micro level: the level of individual human actors, localised tasks, and immediate end-user interactions. Examples: the working conditions of data annotators, the decision-making autonomy of local clinicians, or individual users' interactions with chatbots.¹³²

L2) Meso level: the level of organisational bodies, supply-chains, and industries. Examples: data centre operations, hospital administrative workflows, regional energy grid utilities, or labour markets.¹³³

L3) Macro level: the level of global biophysical systems, geopolitical trading structures, and large-scale societal institutions. Examples: global carbon budgets, biospheric carrying capacities, international mineral supply chains, and democratic public spheres.¹³⁴

Taken together, the life-cycle stages, orders of impact, and levels of analysis constitute the *formal core* of the LCA framework by means of which auditors may reliably diagnose an AI application's systemic impact. To be sure, such impacts are rarely uniform or unequivocal across an application's life-cycle stages and the two axes. A comprehensive diagnosis must therefore also account for structural tensions and trade-offs, where, for example, positive impacts on one axis may coincide with, or actively drive, negative impacts on another.¹³⁵

¹²⁶ See Chouksey et al. 2026; Falk et al. 2026; Orrù 2026; van Wynsberghe 2023

¹²⁷ Chouksey et al. 2026; Crawford and Joler 2018; Dauvergne 2020; Falk et al. 2026; Orrù 2026; van Wynsberghe 2021

¹²⁸ Brevini 2023; Chouksey et al. 2026; Falk et al. 2026

¹²⁹ Vinuesa et al. 2020; Palomares et al. 2021; Onder 2026

¹³⁰ Sætra 2021; Robbins and van Wynsberghe 2022; Waelen and van Wynsberghe 2025

¹³¹ Wang and Blok 2026

¹³² Crawford 2021; Mensah and van Wynsberghe 2025

¹³³ Onder 2026; Robbins and van Wynsberghe 2022; Waelen and van Wynsberghe 2025

¹³⁴ Coeckelbergh and Sætra 2023; van Wynsberghe 2021

¹³⁵ Chouksey et al. 2026; Falk et al. 2026; Sætra 2021; Redin and Ames 2026

3.4.2 Life-Cycle Assessment 2: Sphere-Specific Diagnostics

Within the overarching LCA framework, a systemic approach must employ specific diagnostic methods for the ecological, social, and economic spheres. Specifically, to properly diagnose an AI application's impact within each sphere, auditors must invoke distinct *impact categories*:

Ecological sphere: categories invoked here include global warming potential (GWP), abiotic resource depletion potentials (ADPs), water footprint and water stress, toxicity metrics, eutrophication, acidification, and land use.¹³⁶

Social sphere: categories invoked here include human rights, non-discrimination, labour rights, user autonomy, privacy, power (a-)symmetries, and social cohesion.¹³⁷

Economic sphere: categories invoked here include market diversity, capital concentration, distribution effects, structural labour shifts, and circular economy (CE) metrics.¹³⁸

It is by means of such sphere-specific categories and corresponding scientific methods that auditors can measure the systemic impact of an AI application across its life-cycle stages and along the formal axes of impact orders and levels.

To be sure, it is no easy task to combine all of these parameters, which are based on such diverse categories and measured using such diverse scientific methods—spanning environmental engineering, empirical socio-technical auditing, political economy, and qualitative psychological interviews—into a single, overarching cradle-to-grave impact assessment. Doing so is a difficult but indispensable interdisciplinary undertaking. Only on the basis of such a comprehensive diagnosis of an AI application's systemic impact, and by invoking the implicit norms that constitute the system and its constituent processes, can auditors evaluate whether an AI application

is ethically good or bad, that is, whether it makes a positive or negative contribution to the functioning of the system.

3.4.3 Advantage and Two Worries

The primary advantage of the comprehensive LCA framework is that it grounds AI certification in the concrete reality of complex eco-socio-economic systems. With this framework, the systems paradigm overcomes the problems associated with the principles paradigm's translation assumption, principally those of underdetermination, normative displacement, and limited toolbox mentality.

That being said, the diagnosis assumption also faces serious worries, not least the worry of *system unboundedness*, already familiar from the mapping assumption: namely, the risk of infinite regress when tracing upstream supply-chain dependencies or downstream effects and feedback loops. Since this worry has already been treated at length above (see sec. 3.3.3), it need not be elaborated again here. Beyond unboundedness, two further worries confront the diagnosis assumption:

1) **Framework shopping (again):** The diagnosis assumption faces one of the same worries as the principles paradigm's translation assumption, namely, the proliferation of competing LCA frameworks, background databases, and impact metrics. The problem here is that audits performed using different frameworks may lead to divergent or incompatible impact assessments. In the absence of universally standardised LCA protocols for AI applications, developers and auditing bodies can strategically select or 'shop' for whichever LCA framework yields the most favorable compliance score, thereby once again enabling a sophisticated form of ethics washing under the guise of rigorous assessment.¹³⁹

2) **Primary data deficit:** Another worry is that LCA-based diagnosis is severely constrained by the pervasive data deficit that results from corporate secrecy across the tech sector.¹⁴⁰ The empirical data required to conduct an accurate cradle-to-grave impact assessment—including bill of materials (BOM) for specialised GPUs, power usage

¹³⁶ Chouksey et al. 2026; Falk et al. 2026; ISO 2006a; ISO 2006b; ITU 2014; Röckström et al. 2009

¹³⁷ Ashton 2024, ch. 6; Crawford 2021; Fuchs 2014; Mensah and van Wynsberghe 2025; UNEP 2009; UNESCO 2021; Waelen 2022; Waelen and van Wynsberghe 2025

¹³⁸ Chouksey et al. 2026; Raworth 2012; Sætra 2021; UNEP 2009; Waelen and van Wynsberghe 2025. See also the contributions in Mazzi and Floridi 2023

¹³⁹ See Floridi 2019

¹⁴⁰ Brevini 2023; Chouksey et al. 2026; Falk et al. 2026

effectiveness (PUE) figures for data centres, and the composition of training data sets—are often guarded as trade secrets by major AI developers and hardware manufacturers. Consequently, LCAs of AI applications are currently forced to rely heavily on secondary proxy estimates, scaled commercial baselines, or unverified corporate self-reporting. This opacity obscures the true environmental and human costs of AI applications, leaving auditors unable to properly assess their systemic impact.

In response to the first worry, proponents of the systems paradigm can point out that widely accepted international standards for LCA already exist, most notably the ISO 14040, ISO 14044, ITU-T L.1410, and the UNEP's *Guidelines for social and socio-economic LCA*.¹⁴¹ These established standards provide a consensus-based framework for defining system boundaries, life-cycle inventories, and impact characterisations, thereby offering a clear baseline for harmonised systemic auditing.

In response to the second worry, proponents of the systems paradigm may contend that the lack of primary data is not an inherent flaw in the LCA framework. Rather, it serves as a reminder that certification can only operate against the background of binding governmental and intergovernmental regulation and legislation.¹⁴² With the right regulatory frameworks in place to legally compel corporations to disclose primary life-cycle data, this worry could be substantially alleviated.

3.5 Evaluation Assumption

3.5.1 Method of Immanent-Critical Evaluation

The fourth basic assumption of the systems paradigm is that an AI application's systemic impact can be reliably evaluated in light of the complex system's implicit norms. Although the literature on systemic AI ethics, and on Sustainable AI in particular, frequently insists that AI applications should be evaluated in light of their systemic impact, it specifies next to nothing about how exactly such an evaluation is to be carried out, that is, what concrete methodological

procedure it should follow. To fill this methodological void, we propose that a systemic approach to evaluating the systemic impact of AI applications is best understood as operating by means of a method adopted from critical social philosophy, known as *immanent critique*.

The general idea of immanent critique is to “measure what is being criticised against its own presupposition, against its own law of form”,¹⁴³ that is, against its own implicit norms.¹⁴⁴ This is precisely the driving idea behind systemic impact evaluation: to evaluate the impact of an AI application on the complex system in which it is embedded in light of the system's implicit norms of good functioning. Rather than measure an AI application against abstractly formulated ethical principles applied in an external, top-down manner, immanent critique evaluates whether the application, as part of a complex system across its entire life cycle, contributes positively or negatively to the functioning of the host system, that is, whether it supports and enhances, or disrupts and degrades, the successful reproduction of the system.

This immanent-critical evaluation procedure relies directly on the inputs from the two procedures elaborated in the preceding sections. First, it relies on an understanding of the *system's functioning and implicit norms*, which systems cartography makes explicit (see sec. 3.3). Second, it relies on an understanding of the *AI application's impact on the system's functioning*, which is diagnosed by means of the LCA framework (see sec. 3.4). With these two inputs in hand, the auditors can proceed to evaluate the application's systemic impact as ethically good or bad, that is, as contributing positively or negatively to the functioning and long-term viability of the host system. We can illustrate this procedure with the following toy model:

¹⁴¹ ISO 2006a; ISO 2006b; ITU 2014; UNEP 2009

¹⁴² Maclure and Morin-Martel 2024. See also Corrêa and Mönig 2026b

¹⁴³ Adorno 2018, p. 32 (translation modified, V.W.)

¹⁴⁴ For more on the method of immanent critique, see Jaeggi 2018, ch. 6; Schauer 2023, pp. 338–344; Stahl 2013a; Stahl 2013b

Evaluation of an AI precision-farming application's impact on an agricultural watershed¹⁴⁵

Systems cartography explicates that the implicit norms of an agricultural watershed include ecological integrity, soil fertility retention, hydrological balance, and farmer livelihood security. LCA diagnosis shows that an AI precision-irrigation and crop-yield optimisation application increases short-term commercial output, but drives regional groundwater depletion, relies on hardware supply chains mired in toxic mineral extraction, and triggers third-order rebound dynamics that incentivise monoculture. Measured against the watershed's implicit norms, the AI application represents a *negative contribution*: it optimises the depletion of finite natural resources and pushes the local ecosystem towards biophysical tipping points. Conversely, an AI application configured to operate within established water-table limits and support regenerative soil rotation would represent a *positive contribution* by reinforcing the watershed's ecological integrity.

As the toy model illustrates, evaluating an AI application's impact on the functioning of its host system as good or bad, in light of the latter's implicit norms, has a relatively precise structural meaning. An AI application has a *good impact* across its life cycle insofar as it supports the system in maintaining its internal structure, reproducing the latter's vital ecological, social, and economic conditions, and continuing to fulfill its constitutive purpose. Its impact is even better if the application's life cycle enables the system's constituent processes to fulfill their purposes more effectively than before. Crucially, a good impact also involves keeping multiple pathways of evolution open, preserving the system's capacity to adapt and transform in response to shifting social needs and ecological feedback, and to avoid becoming trapped in path-dependent lock-ins.¹⁴⁶ Conversely, an AI application has a *bad impact* across its life cycle insofar as it disrupts or erodes the system's structural integrity and impairs its successful reproduction. Bad impacts arise when an application's life cycle locks the

system into pathways of development that are unviable or self-undermining in the long term. Examples of this include infrastructural lock-in, resource voracity, structural labour precarity, and systemic rebound effects that undermine the very purpose of the (sub-)systems in which the AI application partakes across its life cycle.¹⁴⁷

3.5.2 Advantage and Two Worries

The advantage of the immanent-critical method is that it grounds the evaluation of an AI application in the concrete implicit norms of the system in which it is embedded, rather than relying on abstract ethical principles applied to isolated artefacts in an external, top-down manner. In doing so, the immanent-critical method allows the systems paradigm to overcome several shortcomings of the principles paradigm, notably the problems of normative underdetermination and displacement, ethics washing, and concealed human and environmental costs.

Despite these advantages, the method of immanent-critical evaluation faces two serious worries:

1) **Collingridge dilemma:** The first worry concerns the temporal mismatch between the deployment and the evaluation of an AI application. Because an immanent-critical evaluation concerns how an AI application actively shapes, supports, or undermines the functioning of complex systems, the application's full ethical-functional impacts can often only be diagnosed after it has already been deployed and become an integral part of the complex system. This exposes immanent-critical evaluation to the so-called Collingridge dilemma:¹⁴⁸ during the early development phase, when an AI application's trajectories are still malleable, its long-term systemic impacts are difficult to foresee; but by the time the application has operated long enough in the wild for its systemic impacts to be empirically measured, the application and its accompanying political and economic interests may already have become structurally locked in and institutionalised.

2) **Functionalist conservatism:** The second worry concerns the potential conservatism of an immanent approach to evaluation.¹⁴⁹ If the evaluative standards we

¹⁴⁵ The toy model draws on insights from Hamdaoui et al. 2024; Kam et al. 2024; Vinod Chandra et al. 2024

¹⁴⁶ Bolte et al. 2022; Bolte and van Wynsberghe 2025; Robbins and van Wynsberghe 2022

¹⁴⁷ Crawford 2021; Robbins and van Wynsberghe 2022

¹⁴⁸ Collingridge 1980; Genus and Stirling 2018

¹⁴⁹ See Jaeggi 2018, ch. 7-8; von Redecker 2018

use are derived from within the host system itself, how can we critically evaluate systems that are themselves fundamentally unjust, ecologically destructive, or structurally exploitative? If an AI application is deployed within a hyper-extractive market economy, a biased judicial system, or an authoritarian surveillance state, evaluating the application based on how well it contributes to the 'successful reproduction' and 'good functioning' of that host system risks legitimising and optimising systemic pathologies. The worry is that, without a set of context-transcending ethical principles, evaluation risks collapsing into a functionalist validation of the status quo, evaluating whether an AI application helps an intrinsically bad system function more efficiently.

Both of these worries present genuine challenges to the immanent-critical method of evaluation. In response to the first worry, proponents of the systems paradigm can point out that there are means available to deal with the temporal gap between early development and long-term systemic visibility, including phased deployment, dynamic auditing, and continuous post-market oversight.¹⁵⁰

In response to the second worry, proponents of the systems paradigm may emphasise that immanent critique is, fundamentally, a transformative methodology rather than a conservative validation of existing practices.¹⁵¹ Grounded in critical social philosophy, immanent-critical evaluation does not merely affirm a host system's explicit self-understanding. Rather, it exposes the operational pathologies and crisis tendencies of a system's structure.¹⁵² A structurally unjust, hyper-extractive, or authoritarian host system is inherently self-undermining and crisis-prone, because its constituent processes and practices tend to undermine its own non-negotiable conditions of reproduction. Consequently, using an AI application to streamline or accelerate an exploitative system would not constitute 'successful reproduction' or 'good functioning'. On the contrary, it would intensify the system's inherent dysfunctionality and accelerate its failure as a sustainable system.

¹⁵⁰ EU 2024; Maclure and Morin-Martel 2024; Pasetti 2026

¹⁵¹ See Celikates 2018; Jaeggi 2018; Stahl 2013a; Stahl 2013b

¹⁵² Jaeggi 2018, ch. 9; von Redecker 2018

4 Conclusion and Outlook

This white paper has examined and problematised the foundational assumptions of two paradigms of AI certification: the *principles paradigm* and the *systems paradigm*. The principles paradigm represents the dominant approach in public policy and industrial standard-setting, seeking to govern AI technologies through top-down, product-centric frameworks that evaluate individual software applications against general ethical principles. In contrast, the systems paradigm offers a structurally minded alternative, shifting the evaluative focus from isolated artefacts to the complex eco-socio-economic systems in which AI applications are embedded throughout their life cycle.

The central insight of this inquiry is that both paradigms rest on a series of basic assumptions, none of which are uncontested, but also not rationally indefensible. The principles paradigm faces serious worries regarding underdetermination, ethics washing, and blindness to environmental and human costs. Yet it can be justified as an administratively manageable, legally predictable, and standardised framework for product-level accountability. Conversely, the systems paradigm faces significant worries concerning system unboundedness, corporate data secrecy, normative projection, and temporal bottlenecks. In turn, it offers a methodology capable of capturing global supply chains, precarious labour conditions, and cumulative environmental impacts.

Despite their divergent philosophical foundations, the two paradigms should not be viewed as irreconcilable dogmas. After all, they are both conceptions of how to assess AI applications in terms of ethical goodness and badness. Beyond this shared goal, there are at least two significant points of connection between the paradigms:

a) Origin of norms: On the surface, the principles paradigm appears to arrive at its principles or norms in a manner *external* to its subject matter, whereas the systems paradigm derives them *internally* from the complex systems in which AI applications are embedded. However, when we examine how the principles paradigm actually formulates its principles—namely, through stakeholder-

informed reflective equilibrium—it becomes clear that it, too, is ultimately concerned with working out the norms that the processes and practices in which we are constantly involved *always already presuppose*. Hence, the difference between the two methods seems to be mainly one of emphasis: whereas immanent critique is explicit about the immanent origin of the norms, reflective equilibrium simply takes it for granted that the norms are already operative in our beliefs and practices.

b) Importance of context: When proponents of the principles paradigm defend their approach against worries about underdetermination of successful translation and conflict arbitration, they consistently fall back on the idea that these issues are settled by the context of an AI application. Yet, adopting methodological contextualism means recognising that an AI model cannot be evaluated in a vacuum, but only relative to the specific purposes, norms, and risk profiles of a concrete deployment domain. In doing so, the principles paradigm already points the way towards a more systemic view of AI applications as technologies that cannot be properly assessed apart from the ecological, social, and economic systems in which they are embedded.

These points of connection indicate the possibility of combining insights from both paradigms in future AI certification schemes. Realising such integrative governance frameworks requires continuous, empirically-informed philosophical reflection, particularly regarding the methodological foundations of AI certification.

It has been a guiding idea of this white paper that any inquiry into the philosophical foundations of AI certification must confront the hard realities of the current world order. As highlighted at the outset, the dominant principles paradigm has developed within a political economy marked by severe asymmetries of political and corporate power. Political and economic actors routinely take advantage of principled certification schemes, exploiting abstract ethical principles, checklists, and stakeholder deliberation procedures as instruments of ethics washing and regulatory capture, publicly signalling virtue while strategically pre-empting, weakening, or delaying binding state regulation.

To be sure, the systems paradigm is not inherently immune to ethics washing and regulatory capture: developers and auditors could attempt to draw convenient system boundaries or selectively interpret implicit norms. However, the systems paradigm explicitly aims at countering these pitfalls. By expanding the audit boundary to encompass cradle-to-grave raw material extraction, labour precarity, energy grid draw, and non-negotiable planetary ceilings, a systemic approach unmask the otherwise hidden costs of AI applications and denies political and corporate actors a 'moral free lunch'.

It remains an essential task for researchers and civil society stakeholders to expose the various ways in which political and economic actors may seek to exploit and co-opt AI certification schemes. Uncovering the weak points in the deep structure of AI ethics and certification that such actors may exploit is a key step towards counteracting ethics washing and regulatory capture, and guarding against similar vulnerabilities in future governance frameworks. Only by continuously bringing to light the hidden assumptions of these frameworks can we chart a viable path forward towards better, more robust, just, and sustainable governance of technologies such as AI.

Acknowledgements

First of all, I want to express my gratitude to the Ministry of Economic Affairs, Industry, Climate Action and Energy of the German State of North Rhine-Westphalia for the financial support to carry out this project, as well as to the University of Bonn for providing the institutional setting and academic freedom necessary for its execution.

I want to thank everyone who has supported me throughout the research and writing process. My sincere thanks go to all of my colleagues at the Center for Science and Thought (CST). Particular thanks go to Chelsea Haramia for insightful discussions during the drafting phase. I am also deeply grateful to my friends Marius Bartmann, Larissa Bolte, Nicklas Kurzböck, and Roman Wagner for numerous discussions in which I had the opportunity to develop and refine the ideas presented herein. Furthermore, I would like to thank the participants of the conferences "AI and the Self" (2-3 June 2026) and "Normativity of AI" (28-29 July 2026) for their stimulating contributions and discussions, which helped me develop the basic structure of this paper. I also want to thank the participants of the joint CST/IWE colloquium on 24 September 2026, where I had the privilege of presenting a draft version of the white paper. Finally, I want to thank my partner, Evangelia Siopi, for her love, patience, and support.

5 References

- [1] Adewale, M. D., Azeta, A., Abayomi-Alli, A., and Sambo-Magaji, A. (2024). "Impact of artificial intelligence adoption on students' academic performance in open and distance learning: A systematic literature review". *Heliyon* 10:e40025, pp. 1-19.
- [2] Adorno, T. W. (2018). *An Introduction to Dialectics (1958)*. Transl. N. Walker. Cambridge: Polity Press.
- [3] AI Ethics Impact Group (2020). "From principles to practice—an interdisciplinary framework to operationalize AI ethics". URL: <https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf> (last accessed on 23/09/2026).
- [4] Ashraf, A. A., Rai, S., Alva, S., Alva, P. D., and Naresh, S. (2025). "Revolutionizing clinical laboratories: The impact of artificial intelligence in diagnostics and patient care". *Diagnostic Microbiology and Infectious Disease* 111:116728, pp. 1-12.
- [5] Ashton, M. F. (2024). *Materials and Sustainable Development*. Oxford: Butterworth-Heinemann.
- [6] Asimov, I. (1950). *I, Robot*. New York: Gnome Press.
- [7] Ayari, H., Guetari, P. R., and Kraïem, P. N. (2026). "Machine learning powered financial credit scoring: a systematic literature review". *Artificial Intelligence Review* 59:13, pp. 1-54.
- [8] Bajwa, J., Munir, U., Nori, A., and Williams, B. (2021). "Artificial intelligence in healthcare: transforming the practice of medicine". *Future Healthcare Journal* 8, pp. e188-e194.
- [9] Baker-Brunnbauer, J. (2021). "TALL Framework for Trustworthy AI Systems". *ROBONOMICS: The Journal of the Automated Economy* 2:17, pp. 1-12.
- [10] Bamgboye, T. T. and Avellán, T. (2025). "Conceptual Map as a Tool for Evaluation in Complexity Science: Usage and Limitations". *Technology, Knowledge and Learning*, URL: <https://link.springer.com/article/10.1007/s10758-025-09846-6>.
- [11] Barøe, K. (2014). "Translational ethics: an analytical framework of translational movements between theory and practice and a sketch of a comprehensive approach". *BMC Medical Ethics* 15:71, pp. 1-7.
- [12] Beauchamp, T. L. and Childress, J. F. (2019). *Principles of Biomedical Ethics*. 8th Edition. Oxford/New York: Oxford University Press.
- [13] Blackburn, S. (1984). *Spreading the Word: Groundings in the Philosophy of Language*. Oxford: Clarendon Press.
- [14] Blackburn, S. (1998). *Ruling Passions*. Oxford: Oxford University Press.
- [15] Blum, C. (2023). "Value Pluralism versus Value Monism". *Acta Analytica* 38, pp. 627-652.
- [16] Boehm, M. (2021). "Hume's 'projectivism' explained". *Synthese* 199, pp. 815-833.
- [17] Bolte, L., Vandemeulebroucke, T., and van Wynsberghe, A. (2022). "From an Ethics of Carefulness to an Ethics of Desirability: Going Beyond Current Ethics Approaches to Sustainable AI". *Sustainability* 14:4472, pp. 1-13.
- [18] Bolte, L. and van Wynsberghe, A. (2025). "Sustainable AI and the third wave of AI ethics: a structural turn". *AI and Ethics* 5, pp. 1733-1742.
- [19] Brevini, B. (2023). "Artificial Intelligence, Artificial Solutions: Placing the Climate Emergency at the Center of AI Developments", in: H. S. Sætra (ed.), *Technology and Sustainable Development: The Promise and Pitfalls of Techno-Solutionism*. Milton Park/Now York: Routledge, pp. 23-33.
- [20] Brink, D. (1984). "Moral realism and the skeptical arguments from disagreement and queerness". *Australasian Journal of Philosophy* 62, pp. 111-125.
- [21] Brogle, K., Kallina, E., Sargeant, H., Shankar, V., Weller, A., Casovan, A., and Bhatt, U. (2025). "Context-specific certification of AI systems: a pilot in the financial industry". *AI Ethics* 5, pp. 4223-4240.
- [22] Castellani, B., Schimpf, C., Wistow, J., Caden, C., Agarwal, J., and Barbrook-Johnson, P. (2025). "Case-based systems mapping: advancing a multimethod approach to social complexity". *International Journal of Social Research Methodology*, pp. 1-21.
- [23] Cath, Y. (2016). "Reflective Equilibrium", in: H. Cappelen, T. S. Gendler, and J. Hawthorne (eds.), *The Oxford Handbook of Philosophical Methodology*. Oxford: Oxford University Press, 213-230.
- [24] Celikates, R. (2018). *Critique as Social Practice: Critical Theory and Social Self-Understanding*. Transl. N. van Steenberg. London/New York: Rowman & Littlefield.
- [25] Chouksey, A., Rajan, A. K., Gurjar, V., Tiwari, R., and Mishra, P. K. (2026). "The Green Paradox: The Climate, Environmental, and Sustainability Implications of Artificial Intelligence". *Global Environmental Change Advances* 6:100029, pp. 1-9.
- [26] Clouser, K. D. and Gert, B. (1990). "A Critique of Principlism". *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 15(2), pp. 219-236.
- [27] Coeckelbergh, M. (2020). *AI Ethics*. Cambridge MA/London: MIT Press.
- [28] Coeckelbergh, M. (2022). *The Political Philosophy of AI. An Introduction*. Cambridge: Polity Press.
- [29] Coeckelbergh, M. and Sætra, H. S. (2023). "Climate change and the political pathways of AI: The technocracy-democracy dilemma in light of artificial intelligence and human agency". *Technology in Society* 75:102406, pp. 1-7.
- [30] Collingridge, D. (1980). *The Social Control of Technology*. London: Frances Pinter Publishers Limited.
- [31] Corrêa, N. K. (2026). *Normative Ethics, Artificial Intelligence, and Value Alignment: From Normative Theory to Engineering Practices*. Cham: Springer.
- [32] Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., and de Oliveira, N. (2023). "Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance". *Patterns* 4:100857, pp. 1-14.
- [33] Corrêa, N. K., Santos, J. W., de Almeida, E., Pasetti, M., Schiavon, D., Panizzon, M., and de Oliveira, N. (2025a). "Codes of ethics in IT: do they work in isolation?" *AI and Ethics* 5, pp. 4337-4349.
- [34] Corrêa, N. K., Schiavon, D., Santos, J. W., Naqvi, F., Hossain, R., Galvão, C., Pasetti, M., and de Oliveira, N. (2025b). "Crossing the principle-practice gap in AI ethics with ethical problem-solving". *AI and Ethics* 5, pp. 1271-1288.
- [35] Corrêa, N. K. and Mönig, J. M. (2024). "Catalog of General Ethical Requirements for AI Certification". URL: <https://arxiv.org/pdf/2408.12289> (last accessed on 22/09/2026)
- [36] Corrêa, N. K. and Mönig, J. M. (2026a). "Catalog of Field-Specific Ethical Requirements for AI Certification". Forthcoming.

Philosophical Foundations of AI Certification

- [37] Corrêa, N. K. and Möning, J. M. (eds.) (2026b). *Standardization of AI Ethics: Stakeholders, Values and Profit*. Bielefeld: Transcript Verlag.
- [38] Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven/London: Yale University Press.
- [39] Crawford, K. and Joler, V. (2018). "Anatomy of an AI System". URL: <https://anatomyofai/> (last accessed on 22/09/2026).
- [40] Cremers, A. B., Englander, A., Gabriel, M., Hecker, D., Mock, M., Poretschkin, M., Rosenzweig, J., Rostalski, F., Sicking, J., Volmer, J., Voosholz, J., Voss, A. and Wrobel, S. (2019). "Trustworthy Use of Artificial Intelligence: Priorities from a Philosophical, Ethical, Legal, and Technological Viewpoint as a Basis for Certification of Artificial Intelligence". URL: <https://publica.fraunhofer.de/entities/publication/13ba8da0-de3c-49aa-ad81-5aca53feafdf> (last accessed on 22/09/2026).
- [41] Davidson, D. (2001). *Inquiries into Truth and Interpretation*. 2nd Edition. Oxford: Clarendon Press.
- [42] Dauvergne, P. (2020). *AI in the Wild: Sustainability in the Age of Artificial Intelligence*. Cambridge MA/London: MIT Press.
- [43] D'Acquisto, G. (2020). "On conflicts between ethical and logical principles in artificial intelligence". *AI & Society* 35, pp. 895-900.
- [44] DePaul, M. R. (1993). *Balance and Refinement: Beyond Coherence Methods of Moral Inquiry*. London/New York: Routledge.
- [45] Digital Catapult (2021). "Ethics Framework". URL: https://www.digicatapult.org.uk/wp-content/uploads/2023/06/DC_AI_Ethics_Framework-2021.pdf (last accessed on 23/09/2026).
- [46] Dong, L., Tang, X., and Wang, X. (2025). "Examining the effect of artificial intelligence in relation to students' academic achievement: A meta-analysis". *Computers and Education: Artificial Intelligence* 8:100400, pp. 1-10.
- [47] Dubljević, V. and Racine, E. (2014). "The ADC of Moral Judgment: Opening the Black Box of Moral Intuitions With Heuristics About Agents, Deeds, and Consequences". *AJOB Neuroscience* 5(4), pp. 3-20.
- [48] European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) (2019). "Ethics guidelines for trustworthy AI". URL: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (last accessed on 22/09/2026).
- [49] European Parliament and the Council (2024). "Regulation (EU) 2024/1689 (Artificial Intelligence Act)". URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng> (last accessed on 22/09/2026).
- [50] Fahim, Y. A., Hasani, I. W., Kabba, S., and Ragab, W. M. (2025). "Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives". *European Journal of Medicine* 30:848, pp. 1-21.
- [51] Falk, S., Ekchajzer, D., Pirson, T., Lees-Perasso, E., Wattiez, A., Biber-Freudenberger, L., Luccioni, S., and van Wynsberghe, A. (2026). "More than carbon: Cradle-to-grave environmental impacts of GenAI training on the Nvidia A100 GPU". *Environmental Impact Assessment Review* 121:108525, pp. 1-16.
- [52] Falk, S. and van Wynsberghe, A. (2024). "Challenging AI for Sustainability: what ought it mean?" *AI and Ethics* 4, pp. 1345-1355.
- [53] Farber, S. (2026). "Machines of justice: A systematic review of AI applications in policing and criminal justice". *The Police Journal: Theory, Practice and Principles*, URL: <https://journals.sagepub.com/doi/10.1177/0032258X261439572> (last accessed on 27/09/2026).
- [54] Feenberg, A. (1999). *Questioning Technology*. New York/London: Routledge.
- [55] Feenberg, A. (2017). *Technosystem: The Social Life of Reason*. Cambridge MA/London: Harvard University Press.
- [56] Floridi, L. (2019). "Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical". *Philosophy & Technology* 32, pp. 185-193.
- [57] Floridi, L. (ed.) (2021). *Ethics, Governance, and Policies in Artificial Intelligence*. Cham: Springer.
- [58] Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., and Vayena, E. (2021). "An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", in: L. Floridi (ed.), *Ethics, Governance, and Policies in Artificial Intelligence*. Cham: Springer, pp. 19-40.
- [59] Floridi, L. and Cows, J. (2019). "A Unified Framework of Five Principles for AI in Society". *Harvard Data Science Review* 1:1, pp. 1-14.
- [60] Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Sri Kumar, M. (2020). "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI". *Berkman Klein Center for Internet & Society*. URL: <http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420> (last accessed on 23/09/2026).
- [61] Fuchs, C. (2014). *Digital Labour and Karl Marx*. New York/London: Routledge.
- [62] Garner, R. T. (1990). "On the genuine queerness of moral properties and facts". *Australasian Journal of Philosophy* 68, pp. 137-146.
- [63] Gauker, C. (1986). "The Principle of Charity". *Synthese* 69, pp. 1-25.
- [64] Genovesi, S. and Möning, J. M. (2022). "Acknowledging Sustainability in the Framework of Ethical Certification for AI". *Sustainability* 14:4157, pp. 1-10.
- [65] Genovesi, S., Möning, J. M., Schmitz, A., Poretschkin, M., Akila, M., Kahdan, M., Kleiner, R., Krieger, L., and Zimmermann, A. (2024). "Standardizing fairness-evaluation procedures: interdisciplinary insights on machine learning algorithms in creditworthiness assessments for small personal loans". *AI and Ethics* 4, pp. 537-553.
- [66] Genus, A. and Stirling, A. (2018). "Collingridge and the dilemma of control: Towards responsible and accountable innovation". *Research Policy* 47, pp. 61-69.
- [67] Gomes, O. and Gubareva, M. (2011). "Complex Systems in Economics and Where to Find Them". *Journal of Systems Science and Complexity* 34, pp. 314-338.
- [68] Google PAIR (2019; third edition updated April 2025). "People + AI Guidebook". URL: <https://pair.withgoogle.com/guidebook/> (last accessed on 23/09/2026).
- [69] Goyal, K., Garg, M., and Malik, S. (2025). "Adoption of artificial intelligence-based credit risk assessment and fraud detection in the banking services: a hybrid approach (SEM-ANN)". *Future Business Journal* 11:44, pp. 1-20.
- [70] Graham, M., Hjorth, I., and Lehdonvirta, V. (2017). "Digital labour and development: impacts of global digital labour platforms and the gig economy on worker livelihoods". *Transfer: European Review of Labour and Research* 23(2), pp. 135-162.
- [71] Griffin, T. A., Green, B. P., and Welie, J. V. (2025). "The ethical wisdom of AI developers". *AI Ethics* 5, pp. 1087-1097.
- [72] Gros, C., Kester, L., Martens, M., and Werkhoven, P. (2025). "Addressing ethical challenges in automated vehicles: Bridging the gap with hybrid AI and augmented utilitarianism". *AI Ethics* 5, pp. 2757-2770.
- [73] Hagendorff, T. (2020). "The Ethics of AI Ethics: An Evaluation of Guidelines". *Minds and Machines* 30, pp. 99-120.
- [74] Halsband, A. (2022). "Sustainable AI and Intergenerational Justice". *Sustainability* 14:3922, pp. 1-11.

- [75] Hamdaoui, H., Hsana, Y., Al kaddouri, H., and Kouddane, N.-E. (2024). "Revolutionizing Agriculture: A Comprehensive Review of AI-Enabled precision Irrigation and Water Quality Forecasting". *Basrah Journal of Agricultural Sciences* 37(2), pp. 354-380.
- [76] Harris, C. and Dubljević, V. (2025). "Navigating the Ethics of Artificial Intelligence". *Encyclopedia* 5:201, pp. 1-25.
- [77] Heathwood, C. (2015). "Monism and Pluralism about Value", in: I. Hirose and J. Olson (eds.), *The Oxford Handbook of Value Theory*. Oxford: Oxford University Press, pp. 136-157.
- [78] Herzog, C. and Blank, S. (2024). "A systemic perspective on bridging the principles-to-practice gap in creating ethical artificial intelligence solutions – a critique of dominant narratives and proposal for a collaborative way forward". *Journal of Responsible Innovation* 11:2431350, pp. 1-23.
- [79] Hoffman, C. H. and Bansal, D. (eds.) (2025). *AI Ethics in Practice: Navigating Academic Insight, Managerial Expertise, and Philosophical Inquiry*. Cham: Springer.
- [80] Hooker, C. (ed.) (2011). *Philosophy of Complex Systems*. Amsterdam: Elsevier.
- [81] Hume, D. (1998). *An Enquiry concerning the Principles of Morals*. Oxford: Oxford University Press.
- [82] Hume, D. (2011). *A Treatise of Human Nature*. Oxford: Oxford University Press.
- [83] Institute of Electrical and Electronics Engineers (IEEE) (2019). "Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, (EADv2)". URL: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf (last accessed on 23/09/2026).
- [84] International Organization for Standardization (ISO) (2006a). "ISO 14040:2006 Environmental management – Life cycle assessment – Principles and framework". URL: <https://www.iso.org/standard/37456.html> (last accessed on 27/09/2026).
- [85] International Organization for Standardization (ISO) (2006b). "ISO 14044:2006 Environmental management – Life cycle assessment – Requirements and guidelines". URL: <https://www.iso.org/standard/38498.html> (last accessed on 27/09/2026).
- [86] International Organization for Standardization (ISO) (2023). "ISO/IEC 42001:2023 Information technology – Artificial intelligence – Management system". URL: <https://www.iso.org/standard/42001> (last accessed on 28/09/2026).
- [87] International Telecommunication Union (ITU) (2014). "Recommendation ITU-T L.1410 (11/2024) – Methodology for environmental life cycle assessments of information and communication technology goods, networks and services". URL: <https://www.itu.int/epublications/publication/itu-t-l-1410-2024-11-methodology-for-environmental-life-cycle-assessments-of-information-and-communication-technology-goods-networks-and-services> (last accessed on 27/09/2026).
- [88] Jaeggi, R. (2018). *Critique of Forms of Life*. Transl. C. Cronin. Cambridge MA/London: The Belknap Press of Harvard University Press.
- [89] Jobin, A., Ienca, M., and Vayena, E. (2019). "The global landscape of AI ethics guidelines". *Nature Machine Intelligence* 1, pp. 389-399.
- [90] Jørgensen, S. E., Patten, B. C., and Straškraba, M. (1992). "Ecosystems emerging: Toward an ecology of complex systems in a complex future". *Ecological Modelling* 62, pp. 1-28.
- [91] Kam, S., Kotecha, R., and Pandey, R. K. (2024). "Towards Sustainable Farming: Leveraging AIoT for Precision Water Management and Crop Yield Optimization". *Procedia Computer Science* 233, pp. 772-781.
- [92] Khakurel, J., Penzenstadler, B., Porras, J., Knutas, A., and Zhang, W. (2018). "The Rise of Artificial Intelligence under the Lens of Sustainability". *Technologies* 6:100, pp. 1-18.
- [93] Kindlydi, I. and Cabral, T. S. (2021). "Sustainability of AI: The Case of the Provision of Information to Consumers". *Sustainability* 13:12064, pp. 1-14.
- [94] Kong, L. and Chang, Y. (2026). "Displacement, Dilemma, and Exit: The Ethical Stakes of Algorithmic Decision-Making". *Social Sciences* 15:593, pp. 1-23.
- [95] Larsson, S. (2021). "AI in the EU: Ethical Guidelines as a Governance Tool", in: A. Bakardjieva Engelbrekt, K. Leijon, A. Michalski, and L. Oxelheim (eds.), *The European Union and the Technology Shift*. Cham: Palgrave Macmillan, pp. 85-111.
- [96] Lee, M. J. H. (2010). "The problem of 'thick in status, thin in content' in Beauchamp and Childress' Principlism". *Journal of Medical Ethics* 36(9), pp. 525-528.
- [97] Lemmens, P., Blok, V., and Zwier, J. (2017). "Toward a Terrestrial Turn in Philosophy of Technology". *Techné: Research in Philosophy and Technology* 212(2/3), pp. 114-126.
- [98] Lewis, D., Hogan, L., Filip, D., and Wall, P. J. (2020). "Global Challenges in the Standardization of Ethics for Trustworthy AI". *Journal of ICT Standardization* 8(2), pp. 123-150.
- [99] Lindgren, S. (2023a). *Critical Theory of AI*. Cambridge: Polity Press.
- [100] Lindgren, S. (ed.) (2023b). *Handbook of Critical Studies of Artificial Intelligence*. Cheltenham/Northampton: Edward Elgar Publishing.
- [101] Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin.
- [102] Maclure, J. and Morin-Martel, A. (2025). "AI Ethics' Institutional Turn". *Digital Society* 4:18, pp. 1-25.
- [103] Madanchian, M., Taherdoost, H., and Mohamed, N. (2023). "AI-Based Human Resource Management Tools and Techniques: A Systematic Literature Review". *Procedia Computer Science* 229, pp. 367-377.
- [104] Maner, W. (1999). "Is Computer Ethics Unique?" *Etica & Political/Ethics & Politics* 1(2).
- [105] Marcoccia, C. (2026). "Citizens Infrastructures as a Way to Govern AI's Power to Shape our Shared Representations. How to Make Sure Auditing Institutions Are Aligned with Values and Societal Expectations?" in: N. K. Corrêa and J. M. Mönig (eds.), *Standardization of AI Ethics: Stakeholders, Values and Profit*. Bielefeld: Transcript Verlag, pp. 39-66.
- [106] Max, R. (2026). "Guidelines without guidance: underdetermination in AI ethics". *AI and Ethics* 6:323, pp. 1-10.
- [107] Mazzi, F. and Floridi, L. (eds.). (2023). *The Ethics of Artificial Intelligence for the Sustainable Development Goals*. Cham: Springer.
- [108] McGregor Richmond, K. and Primeau, C. (2026). "AI in the criminal justice system: From 'AI ethics' to 'hermeneutic risk'". *Medicine, Science and the Law* 66(2), pp. 75-79.
- [109] McShane, K. (2004). "Ecosystem Health". *Environmental Ethics* 26:3, pp. 227-245.
- [110] Mensah, M. and van Wynsberghe, A. (2025). "Sustainable AI meets feminist african ethics". *AI and Ethics* 5, pp. 4293-4303.
- [111] Mensch, M., Seeliger, A., and Wellhöfer, J. (2026). "Navigating Challenges and Opportunities of Standards Developing Organizations in Times of Rapid Technological Advancement", in: N. K. Corrêa and J. M. Mönig (eds.), *Standardization of AI Ethics: Stakeholders, Values and Profit*. Bielefeld: Transcript Verlag, pp. 143-161.

Philosophical Foundations of AI Certification

- [112] Microsoft (2022). "Microsoft Responsible AI Standard, v2". URL: <https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Responsible-AI-Standard-General-Requirements.pdf> (last accessed on 23/09/2026).
- [113] Mills, S. and Sætra, H. S. (2025). "Algorithms in the room: AI, representation, and decisions about sustainable Futures". *Technovation* 145:103304, pp. 1-12.
- [114] Mittelstadt, B. (2019). "Principles alone cannot guarantee ethical AI". *Nature Machine Intelligence* 1, pp. 501-507.
- [115] Moor, J. H. (1995). "Is ethics computable?" *Metaphilosophy* 26, pp. 1-21.
- [116] Moor, J. H. (2006). "The nature, importance, and difficulty of machine ethics". *IEEE Intelligent Systems* 21, pp. 18-21.
- [117] Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., and Floridi, L. (2021). "Ethics as a Service: A Pragmatic Operationalisation of AI Ethics". *Minds and Machines* 31, pp. 239-256.
- [118] Munn, L. (2023). "The uselessness of AI ethics". *AI Ethics* 3, pp. 869-877.
- [119] Nakano, M. (2026). "Bridging the AI Ethics Gap: A Tripartite Framework for Accountability, Implementation, and Governance". *International Journal of Software Innovation* 14(1), pp. 1-14.
- [120] Nguyen, D. H., Tran, K. D., Luong, N. A., Luong, M. A., Thomassey, S., and Tran, K. P. (2026). "Artificial Intelligence for Human Resource Management: A Review, Perspective, and Case Study", in: S. Thomassey, K. P. Tran, F. Vanderhaegen, and T.-M. Choi (eds), *Responsible Artificial Intelligence and Data Science (RAIDS 2024)*. Cham: Springer, pp. 124-138.
- [121] Neuhouser, F. (2018). "The Normativity of Forms of Life", in: A. Allen and E. Mendieta, *From Alienation to Forms of Life: The Critical Theory of Rahel Jaeggi*. University Park PA: Pennsylvania State University Press, pp. 59-75.
- [122] Odenbaugh, J. (2011). "Complex Ecological Systems", in: C. Hooker (ed.), *Philosophy of Complex Systems*. Amsterdam: Elsevier, pp. 421-439.
- [123] OECD (2019/2024). "OECD AI Principles". URL: <https://www.oecd.org/en/topics/ai-principles.html> (last accessed on 22/09/2026).
- [124] OECD (2021). "Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems". URL: https://www.oecd.org/content/dam/oecd/en/publications/reports/2021/06/tools-for-trustworthy-ai_0e36bb08/008232ec-en.pdf (last accessed on 23/09/2026). See also the accompanying OECD database "Catalogue of Tools & Metrics for Trustworthy AI": <https://oecd.ai/en/catalogue/tools> (last accessed on 23/09/2026).
- [125] Olson, J. (2014). *Moral Error Theory: History, Critique, Defence*. Oxford: Oxford University Press.
- [126] Onder, O. (2025). "Planetary health in the age of artificial intelligence: a structural-ethical inquiry into sustainable AI in healthcare". *AI & Society* 41, pp. 1967-1978.
- [127] Ong, L. M. and Findlay, M. (2023). "A Realist's Account of AI for SDG's: Power, Inequality and AI in Community", in: F. Mazzi and L. Floridi (eds.), *The Ethics of Artificial Intelligence for the Sustainable Development Goals*. Cham: Springer, pp. 43-62.
- [128] Orrù, E. (2026). "Small data AI as a more sustainable next-wave AI? Exploring a potential sociotechnical shift away from the voracity of big data". *AI & Society* 41, pp. 1035-1051.
- [129] Palomares, I., Martínez-Cámara, E., Montes, R., García-Moral, P., Chiacchio, M., Chiacchio, J., Alonso, S., Melero, F. J., Molina, D., Fernández, B., Moral, S., Marchena, R., de Vargas, J. P., and Herrera, F. (2021). "A panoramic view and SWOT analysis of artificial intelligence for achieving the sustainable development goals by 2030: progress and prospects". *Applied Intelligence* 51, pp. 6497-6527.
- [130] Pasetti, M. (2026). "Should AI Auditors Be Audited? Challenges and Paths for Meta-Auditing Artificial Intelligence", in: N. K. Corrêa and J. M. Mönig (eds.), *Standardization of AI Ethics: Stakeholders, Values and Profit*. Bielefeld: Transcript Verlag, pp. 15-38.
- [131] Peylo, C., Slama, D., Hallensleben, S., Hauschke, A., and Hildebrandt, S. (2022). "VCIO based description of systems for AI trustworthiness characterisation". URL: https://www.vde.com/resource/blob/2242194/a24b13db01773747e6b7bba4ce20ea60/vcio-based-description-of-systems-for-ai-trustworthiness-characterisationvde-spec-90012-v1-0--en-data.pdf?trk=public_post_comment-text (last accessed on 23/09/2026).
- [132] Plociennik, C., Watjanatepin, p., Van Acker, K., and Ruskowski, M. (2025). "Life Cycle Assessment of Artificial Intelligence Applications: Research Gaps and Opportunities". *Procedia CIRP* 135, pp. 924-929.
- [133] Polk, M., Palmer, H., and Raviola, E. (2026). "Unbounded problems in a bounded world: using the core features of wicked problems to understand and manage the dilemmas that arise in cross-sector collaborations around complex societal problems". *Sustainability Science*, URL: <https://link.springer.com/article/10.1007/s11625-026-01877-x> (last accessed on 27/09/2026).
- [134] Pranavi, E., Mahindrakar, S., Latha, T. M., Lakshmi, V. V., and Mamatha, J. (2026). "Artificial Intelligence Reshaping Human Resource Management", in: A. Kumar and S. Mozar (eds), *Proceedings of the 7th International Conference on Communications and Cyber Physical Engineering (ICCCE 2024)*. Cham: Springer, pp. 1144-1152.
- [135] Ramachandran, M. (2026). *Engineering Ethics of AI by Design: Principles, Practices, and Frameworks for Responsible Artificial Intelligence*. Cham: Springer.
- [136] Rawls, J. (1999). *A Theory of Justice. Revised Edition*. Cambridge MA: The Belknap Press of Harvard University Press.
- [137] Rawls, J. (2003). *Justice as Fairness. A Restatement*. Cambridge MA: The Belknap Press of Harvard University Press.
- [138] Raworth, J. (2012). "A Safe and Just Space for Humanity: Can We Live within the Doughnut?" *Oxfam Discussion Papers*, URL: https://www-cdn.oxfam.org/s3fs-public/file_attachments/dp-a-safe-and-just-space-for-humanity-130212-en_5.pdf (last accessed on 27/09/2026).
- [139] Redin, D. M. and Ames, M. C. (2026). "Beyond Greenwashing: Governing Sustainable AI for the Common Good Through a Virtue Ethics and Human-Centered AI Framework". *Journal of Business Ethics*, pp. 1-23.
- [140] Robbins, S. and van Wynsberghe, A. (2022). "Our New Artificial Intelligence Infrastructure: Becoming Locked into an Unsustainable Future". *Sustainability* 14:4829, pp. 1-11.
- [141] Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S., Lambin, E., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P. K., Costanza, R., Svedin, U., ... and Foley, J. (2009). "Planetary Boundaries: Exploring the Safe Operating Space for Humanity". *Ecology and Society* 14(2), URL: <http://www.jstor.org/stable/26268316> (last accessed on 27/09/2026).
- [142] Scanlon, T. M. (2003). "Rawls on Justification", in: S. Freeman (ed.), *The Cambridge Companion to Rawls*. Cambridge: Cambridge University Press, pp. 139-167.
- [143] Scanlon, T. M. (2014). *Being Realistic About Reasons*. Oxford: Oxford University Press.
- [144] Schauer, A. (2023). *Mensch ohne Welt. Eine Soziologie spätmoderner Vergesellschaftung*. Berlin: Suhrkamp Verlag.
- [145] Schmidt, M. W. (2024). "Defining the method of reflective equilibrium". *Synthese* 203:170, pp. 1-23.

- [146] Shuju, F. and Aifen, X. (2023). "International Soft Law Governance of Artificial Intelligence: Advantages, Approaches, and Credibility". *Journal of Business Theory and Practice* 11(3), pp. 35-45.
- [147] Stahl, T. (2013a). *Immanente Kritik: Elemente einer Theorie sozialer Praktiken*. Frankfurt/New York: Campus Verlag.
- [148] Stahl, T. (2013b). "What is Immanent Critique?" *SSRN Working Papers*, URL: <http://ssrn.com/abstract=2357957> (last accessed on 28/09/2026).
- [149] Stellinga, L., Korenhof, P., and Blok, V. (2025). "Bio-Centered Artificial Intelligence: Towards a Progressive AI-biosphere Relation Through the Concepts of Poiesis and Mimesis". *Topoi*, pp. 1-11.
- [150] Stilgoe, J. (2019). *Who's Driving Innovation? New Technologies and the Collaborative State*. Cham: Palgrave Macmillan
- [151] Sætra, H. S. (2021). "AI in Context and the Sustainable Development Goals: Factoring in the Unsustainability of the Sociotechnical System". *Sustainability* 13: 1738, pp. 1-19.
- [152] Sætra, H. S. and Rogerson, S. (2026). "Four visions on the future of global governance, AI, and climate change". *AI & Society* 41, pp. 4751-4766.
- [153] Taddeo, M. and Floridi, L. (2018), "How AI can be a force for good". *Science* 361: 6404, pp. 751-752.
- [154] Bosschaert, M. T. and Blok, V. (2026). "Towards an Integrated Framework in the Philosophy of Technology". *Techné: Research in Philosophy and Technology* 30, pp. 29-56.
- [155] Thurner, S., Klimek, P., and Hanel, R. (2018). *Introduction to the Theory of Complex Systems*. Oxford: Oxford University Press.
- [156] Tranquillo, J. (2019). *An Introduction to Complex Systems*. Cham: Springer.
- [157] Úbeda-García, M., Marco-Lajara, B., Zaragoza-Sáez, P. C., and Poveda-Pareja, E. (2025). "Artificial intelligence, knowledge and human resource management: A systematic literature review of theoretical tensions and strategic implications". *Journal of Innovation & Knowledge* 10:100809, pp. 1-12.
- [158] United Nations Environment Programme (UNEP) (2009). "Guidelines for Social Life Cycle Assessment of Products". URL: <https://wedocs.unep.org/rest/api/core/bitstreams/a8c00a35-628a-4dce-b8ac-ceec36e481a1/content> (last accessed on 28/09/2026).
- [159] United Nations Educational, Scientific and Cultural Organization (UNESCO) (2021). "Recommendation on the Ethics of Artificial Intelligence". URL: <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics> (last accessed on 22/09/2026).
- [160] van Wynsberghe, A. (2021). "Sustainable AI: AI for sustainability and the sustainability of AI". *AI and Ethics* 1, pp. 213-218.
- [161] van Wynsberghe, A., Vandemeulebroucke, T., Bolte, L., and Nachid, J. (eds.). (2023). *Towards the Sustainability of AI: Multi-Disciplinary Approaches to Investigate the Hidden Costs of AI*. Basel: MDPI Books.
- [162] Verbeek, P.-P. (2021). "The Empirical Turn", in: S. Vallor (ed.), *The Oxford Handbook of Philosophy of Technology*. Oxford: Oxford University Press, pp. 35-54.
- [163] Vieriu, A. M. and Petrea, G. (2025). "The Impact of Artificial Intelligence (AI) on Students' Academic Development". *Education Sciences* 15:343, pp. 1-12.
- [164] Vinod Chandra, S.S., Anand Hareendran, S., and Ghassan Faisal, A. (2024). "Precision farming for sustainability: An agricultural intelligence model". *Computers and Electronics in Agriculture* 226:109386, pp. 1-11.
- [165] Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., and Fuso Nerini, F. (2020). "The Role of Artificial Intelligence in Achieving the Sustainable Development Goals". *Nature Communications* 11:233, pp. 1-10.
- [166] von Redecker, E. (ed.) (2018). "Progress, Normativity, and the Dynamics of Social Change: An Exchange between Rahel Jaeggi and Amy Allen", in: A. Allen and E. Mendieta (eds.), *From Alienation to Forms of Life: The Critical Theory of Rahel Jaeggi*. University Park PA: Pennsylvania State University Press, pp. 156-187.
- [167] Waelen, R. (2022). "Why AI Ethics Is a Critical Theory". *Philosophy and Technology* 35:9, 1-16.
- [168] Waelen, R. and van Wynsberghe, A. (2025). "Considering the Social and Economic Sustainability of AI". *Science and Engineering Ethics* 31:19, pp. 1-13.
- [169] Wang, H. and Blok, V. (2026). "ELSA Labs as Method to Study AI-Based Systems at the Micro, Meso, and Macro Level", in: S. S. Gouveia (ed.), *The Palgrave handbook on the Ethics of Artificial Intelligence*. Cham: Springer, pp. 17-33.
- [170] World Health Organization (WHO) (2021). "Ethics and governance of artificial intelligence for health: WHO guidance". URL: <https://www.who.int/publications/i/item/9789240029200> (last accessed on 22/09/2026).
- [171] Wiener, N. (1988). *The Human Use of Human Beings: Cybernetics and Society*. New York: Grand Central Publishing.
- [172] Zhou, J. and Chen, F. (2023). "AI ethics: from principles to practice". *AI & Society* 38, 2693-2703 (2023).
- [173] Zuhair, M., Kadhim, K., Riyadh Abdulrahman, S. A., Hussan, A. O., Sharad, A., and Voronzhesky, K. (2026). "Artificial Intelligence in Criminal Justice: Examining Legal, Ethical, and Social Impacts", in: Z. Molamohamadi, E. Babaee Tirkolaei, A. Mirzazadeh, and G.-W. Weber (eds), *Optimization and Data Science in Industrial Engineering (ODSIE 2025)*. Cham: Springer, pp. 611-622.