Swiss Journal of Business

Established 1947 as *Die Unternehmung*

Published on behalf of the Schweizerische Gesellschaft für Betriebswirtschaft (SGB)

1/26
Volume 80
ISSN 2944-3741

Nomos

Inlibra

Special Issue

Responsible and Human-Centered Artificial Intelligence

Guest Editors: Stefan Güldenberg, Uta Wilkens, Tom Stoneham

Swiss Journal of Business

1/2026

Organ der Schweizerischen Gesellschaft für Betriebswirtschaft (SGB)
L'organe de la Société suisse de gestion d'entreprise

Volume 80
Pages 1–94

Published on behalf of the Schweizerische Gesellschaft für Betriebswirtschaft (SGB)
Established 1947 as *Die Unternehmung*

Editorial Office: Prof. Dr. Stefan Güldenberg, EHL Hospitality Business School, EHL Campus Lausanne, Route de Berne 301, CH-1000 Lausanne 25, email: stefan.guldenberg@ehl.ch

Contents

Perspective Articles

Editorial to the Special Issue

Responsible and Human-Centered Artificial Intelligence

Standards, Processes and Behaviors

Stefan Güldenberg, Uta Wilkens, Tom Stoneham

With this special issue on Responsible and Human-Centered Artificial Intelligence (AI), we celebrate the start of the 80th anniversary volume of the *Swiss Journal of Business (*Established 1947 as *Die Unternehmung)*. For 80 years the *Swiss Journal of Business* has aimed to disseminate new findings in business research, to draw attention to important problems in society, research and business practice, to set research agendas, to present scientifically sound solutions and generally to promote the exchange between science and practice. With over 350 issues and more than 2,000 published research articles, we are proud that our transition to a new name and full open access is bearing fruit and that we have now reached an audience of close to 150,000 international online readers yearly. We are very much looking forward to further expanding our reach and impact by focusing on grand challenges and interdisciplinary topics that broaden the horizon of business research and practice. The topic of this first special issue of our 80th volume is an excellent example of this.

Many in media, business and politics talk about responsible and human-centered AI. But what exactly does it mean, and why has so little of it been implemented despite the pressing urgency, given the technological speed of AI development and the influence it has on human beings and their workplaces? This special issue is dedicated to the grand challenge of AI. It aims to stimulate the debate on responsible AI in corporate business ethics from an institutional and configurational perspective. The focus is on developing and implementing regulations, declarations and standards to ensure responsible AI design and deployment, and on integrating AI into business operations ethically. Emphasis is also placed on the human elements of AI ethics, including individual decision-making, reflective practices, AI literacy and leadership behaviors.

The application of AI in the workplace and the reflection on ethics related to AI-generated content, solutions and consequences for users and customers, is a matter for all

stakeholders – ethics is always an individual responsibility as well as a corporate one. Core research questions reflect on individual practices in AI-augmented work systems, responsible and reflective user behavior, requirements for maintaining ethical choices, applications, and behavior. This also includes leadership roles in ensuring transparency, fairness and equality in AI-assisted domains or algorithmic management.

The EU AI Act as well as the EU General Data Protection Regulation (GDPR) helps to systemize and understand risks and hidden ethical threats of corporate AI applications. This includes addressing the risk levels outlined in the EU AI Act specifying their criteria for responsible AI in corporate declarations and aligning these frameworks with corporate values and existing reporting systems such as ESG. The core research questions are related to the tension and harmonization between different standards, their impact on mitigating risks, and achieving sustainable development goals (SDGs). Another core area of research involves the visibility, acceptance and bargaining of declarations in the face of power differences.

From the perspective of corporate management and business ethics the responsible use of AI is more than just appealing, reflecting user behavior and non-negotiable fundamental values - it is a corporate responsibility for strategy, competitiveness and systems' resilience. For generating sustainable solutions, enhancing process resilience, and fostering organizational learning while using AI systems, corporate management must create structures, institutionalize governance mechanisms, and foster AI literacy for a responsible deployment of AI – not only on a technical basis for compliance regulation but as a social and cultural value. So far, little is known about organizational strategies, practices and outcomes with respect to responsible AI. Given the potential dislocation of accountability, core research questions address the location of responsibility, characteristics of responsibility-enhancing processes, the effectiveness of norms and practices, as well as consequences for SDGs.

In response to this challenging context, this special issue consists of seven articles, three full-length research articles and four shorter perspective contributions, all of which describe how companies reflect on AI ethics in their corporate strategy. All articles in this special issue share a common aim: to reflect and discuss the relevance and far-reaching implications of responsible and human-centered AI in practice.

The research articles provide conceptual frameworks, empirical evidence and case studies insights on how responsible and human-centered AI can be seen as a process and outcome, through the lens of ordonomics and against a socio-technical framework. They explain in great detail why responsible and human-centered AI should not only be seen as a technological challenge but a cultural and social one that requires smart organizational design and governance, how organizations construct collective interpretations of responsible AI along the dimensions of legitimacy, suitability, necessity, and proportionality, how corporations continuously adapt their practices while seeking for legitimization within a field of collective interpretation of responsible AI, and that even among companies with high obligations in public transparency only a minority systematically applies ethical AI standards on a high level where it is not legally required.

The perspective articles pinpoint selected discourses on AI ethics, broaden the view for future research, and address new topics for business ethics. In doing so, the contributing authors succeed in highlighting the necessity of responsible and human-centered AI as well as its practical implementation from various disciplinary and sector perspectives regarding

standards, processes and behaviors. It quickly becomes evident that the implementation of responsible and human-centered AI has implications for multiple disciplines, each characterized by distinct research traditions and practical relevance, necessitating their adequate representation.

The first research contribution "Responsible Use of Artificial Intelligence as Continuous Proportionalization" by *Simon Sturm, Florian Krause and Benjamin van Giffen* presents findings from original case study research in the fashion industry. The authors conceptualize responsible AI as both a process and outcome of social evaluation and propose a model ("continuous proportionalization") that explains how organizations construct collective interpretations of responsible AI along the dimensions of legitimacy, suitability, necessity, and proportionality. They argue that there is no established understanding of the responsible use of AI; instead corporations continuously adapt their practices while seeking for legitimization within a field of collective interpretation of responsible AI. The conceptual framework is illustrated with the case study of AI-based fashion image generation at OTTO, Germany's largest e-commerce company.

The second research contribution "Artificial Intelligence as a Socio-Economic Dilemma: Ordonomic Diagnosis–Reflection–Design for Education, Work and Governance" by *Patrick Hedfeld* analyzes AI through the lens of ordonomics, a normative-institutional approach that connects economic rationality with ethical reflection. While most discussions in AI ethics focus on principles such as fairness, transparency, and accountability, fewer studies address how these principles can be institutionalized through incentive-compatible rules. This research paper therefore conceptualizes AI not as a primarily technological challenge but as a social order problem that requires institutional design and governance. *Hedfeld* explicitly maps the classical ordonomic three-level schema - actor, institutional order, and market/discourse - onto an applied heuristic of Diagnosis–Reflection–Design, demonstrating how this triad operationalizes ordonomic reasoning for the AI context. Building on this foundation, the paper identifies and categorizes key AI-related social dilemmas (economic, epistemic, ethical, and educational). The analysis develops differentiated responsibilities across levels of coordination and proposes rule-based cooperation solutions that align individual incentives with collective welfare. By linking ordonomics to current frameworks such as Responsible AI, algorithmic accountability, and the EU AI Act, the paper positions ordonomics as a design-oriented ethics that bridges normative ideals and institutional economics. The result of the paper is a framework for diagnosing conflicts, reflecting responsibilities, and designing cooperative solutions that reconcile innovation with social responsibility.

In the third research contribution "Towards AI Governance in DAX40: A Typology of Organizational Guidelines for Self-Regulation" *Niklas Obermann, Daniel Lupp and Uta Wilkens* provide empirical evidence to organizational practices in coping with challenges of ethical AI. The authors systematically evaluate the self-regulating guidelines for AI ethics of the German DAX40 corporations against a socio-technical framework specifying different ways of how to demonstrate responsibility, whether it is directed towards the trustworthiness of the technology or may also include issues of organizational or personnel development. The outcome of their qualitative content analysis is "A typology of organizational guidelines for self-regulation", distinguishing DAX40 corporations with (1) non-codified self-regulation, (2) symbolic-technical self-regulation, and (3) comprehensive socio-technical self-regulation. Only the latter mirrors a broader discourse on AI ethics

including social dimensions of ethics in addition to technological characteristics of AI. The differences indicate that even among companies with high obligations in public transparency only a minority systematically applies ethical standards on a high level where it is not legally binding.

In the first of the four perspective contributions "Trust and Responsibility in AI: An Interdisciplinary Social-Sector Perspective" by *Nathan Chappell* the author states that the rapid adoption of artificial intelligence has intensified debates about responsibility, ethics, and trust. While regulatory frameworks and organizational ethics statements are proliferating, responsible AI is too often treated as compliance or reputation management rather than an organizing principle of practice. His perspective paper argues that social-sector organizations - including nonprofits, NGOs, and other mission-driven institutions - offer an instructive lens for rethinking responsible AI because they operate with structural vulnerability and high trust dependence. By drawing on business ethics, organizational theory, nonprofit and social-sector management, and human flourishing scholarship, *Chappell* proposes shifting from harm-avoidance toward trust-centered, flourishing-oriented AI integration.

The second perspective paper "Responsible und Human-Centered "AI" - Some Ethical Considerations" by *Peter G. Kirchschläger* provides an ethical as well as practical guidance towards responsible and human-centered AI. *Kirchschläger* challenges the term AI by referring to it as "data-based systems (DS)" and sees the necessity to identify ethical opportunities and risks of DS in order to promote the former and in order to avoid the latter – for the benefit of all people and the planet earth. He states in his perspective paper that companies can contribute to the realization of DS with ethics by, first, living up to the exclusive human responsibility for machines; second, while running innovation- and research-processes, by implementing always right from the start an interaction between ethics and technologies; third, by promoting global human rights-based regulation of DS as well as the establishment of an International Data-Based Systems Agency (IDA) at the UN enforcing this global regulation of DS.

The third perspective paper "Defining Knowledge in the Age of Society 5.0" by *Susanne Durst* takes a knowledge management perspective by introducing the concept of responsible knowledge management (rKM) and its usefulness for implementing ethically accepted AI solutions in organizations. Illustrative examples are presented to demonstrate the latter. *Durst* states in her paper that integrating the underlying principles of rKM into discussions related to responsible and human-centred AI in business ethics is expected to lead to the development and execution of more inclusive and responsible solutions to addressing the grand challenges at hand. She argues that this way of thinking can also contribute to achieving the United Nations Sustainable Development Goals, in particular Goal 5 "Gender Equality," Goal 10 "Reduced Inequalities," and finally Goal 17 "Partnerships", and that the use of AI raises the question of digital inequality, which is why a human-centred, inclusive, and collaborative approach such as rKM is more important than ever. The paper concludes with a series of research questions that serve as an outlook and inspiration for further reflection on rKM and ethically acceptable AI solutions in companies.

The fourth and final perspective paper "Intimate Machines, Disturbed Minds: Managing the Affective Cost of AI" by *Leona Chandra Kruse* and *Patrick Mikalef* focuses on the affective costs of AI on an individual as well as team level in organizations. The authors suggest preventive as well as corrective regulation and governance principles that

seek to limit problematic affective dynamics upstream while at the same time respond to relational and emotional disruptions once AI becomes embedded in work practices. At the end of their perspective paper *Chandra Kruse/Mikalef* propose a further research agenda for better managing the affective costs of AI in organizations.

This special issue provides a broad overview of the grand challenges, opportunities and ethical risks of human-centered AI in organizations and society at large. It offers both theoretically grounded and practice-oriented approaches and examples of how ethical governance standards, processes, behaviors, and regulations can be established. It also discusses potential limitations, shows existing contradictions, highlights research gaps as well as future research questions of responsible and human-centered AI. A central theme that emerges across all contributions is the critical importance of going beyond direct technology-related criteria of ethics and taking relational and social aspects into consideration, even if they are indirectly addressed, thereby exemplifying Kranzberg's First Law: "Technology is neither good nor bad; nor is it neutral" (1986, 545). AI ethics is conceptualized as an issue incorporated in all spheres of business, from governance and process design to people management and individual responsibility. The rapid adoption of AI technologies is rushing ahead of both hard and soft regulation, and of cultural and social norms which might guide business leaders, creating both risks and opportunities which need to be bounded by explicit engagement with ethics.

Given the current dynamic and disruptive technological developments we are pretty sure that this special issue is more the beginning than the end of an ethical discourse on responsible and human-centered AI in organizations and society. May it inspire future research and practical actions on ethical governance standards, processes, behaviors, and regulations regarding a more responsible and human-centered use of AI. We would like to thank all the authors involved in this special issue for their insightful contributions. We are especially grateful to our dedicated reviewers, who have made a significant contribution to ensuring the quality of this special issue.

References

Kranzberg, M. (1986). Technology and History: 'Kranzberg's Laws.'*Technology and Culture* 27(3): 544–560. https://doi.org/10.2307/3105385.

Stefan Güldenberg, Prof. Dr., is Managing Editor of the Swiss Journal of Business, President of the Swiss Society for Business and Management and Full Professor as well as Academic Director at the Graduate School of the EHL Hospitality Business School, Lausanne.

Address: EHL Hospitality Business School, HES-SO, University of Applied Sciences and Arts Western Switzerland, Route de Berne 301, 1000 Lausanne, 25, Switzerland,
E-Mail: stefan.guldenberg@ehl.ch, sjb@nomos-journals.de
ORCID: https://orcid.org/0000-0002-8698-4113

Uta Wilkens, Prof. Dr., is Full Professor of Work, Human Resources & Leadership at Ruhr University Bochum, Germany, member of acatech "Learning Systems", Chairperson of the Competence Center HUMAINE and the HUMAINE Network e.V., an Association for Human-Centered AI.

Address: Ruhr University Bochum, Institute of Work Science, Chair of Work, Human Resource & Leadership, Universitätsstr. 150, 44801 Bochum, Germany,
E-Mail: uta.wilkens@rub.de
ORCID: https://orcid.org/0000-0002-7485-4186

Tom Stoneham, Prof. Dr., is Full Professor of Philosophy at the University of York, UK. He is Ethics Lead for the UKRI Centre for Doctoral Training in Safe AI Systems and Convenor of the MA in Applied Ethics and Governance of Data Privacy, and President of the International Berkeley Society.

Address: Department of Philosophy, University of York, Heslington, York, YO10 5DD, UK, E-Mail: tom.stoneham@york.ac.uk
ORCID: https://orcid.org/0000-0001-5490-4927

Received: 01.09.2025 | Revised: 14.12.2025 | Accepted: 16.12.2025

Responsible Use of Artificial Intelligence as Continuous Proportionalization: Fashion Image Generation at OTTO

Simon Sturm, Florian Krause, Benjamin van Giffen

Abstract: The inherent ambivalence of machine learning-based artificial intelligence (AI) technologies makes ensuring their responsible use a pressing concern. While the literature converges on the importance of governance at the organizational level, uncertainty remains about what responsible (use of) AI actually is. We conceptualize responsible AI as both a process and outcome of social evaluation. We propose a model ("continuous proportionalization") that explains how organizations construct collective interpretations of responsible AI along the dimensions of legitimacy, suitability, necessity, and proportionality. We illustrate the model through a case study of AI-based fashion image generation at Germany's largest e-commerce company, OTTO.

Keywords: Artificial Intelligence, Responsible AI, AI Ethics, Image Generation, E-Business, Technology Management, Business Ethics, Corporate Governance

Verantwortungsvolle Nutzung Künstlicher Intelligenz als kontinuierliche Proportionalisierung: Fashion-Bildgenerierung bei OTTO

Zusammenfassung: Die inhärente Ambivalenz und die zunehmende Nutzung künstlicher Intelligenz (KI) werfen drängende normative Fragen auf. Die Literatur ist sich einig, dass verantwortungsvolle KI-Nutzung Governance auf der organisationalen Ebene erfordert. Uneinigkeit besteht jedoch darüber, was verantwortungsvolle KI-Nutzung ist und wie sie entsteht. Dieser Artikel betrachtet verantwortungsvolle KI-Nutzung als Prozess und Ergebnis sozialer Evaluation. Er entwickelt ein konzeptionelles Modell („kontinuierliche Proportionalisierung"), das erklärt, wie Organisationen kollektive Interpretationen verantwortungsvoller KI-Nutzung entlang der Dimensionen Legitimität, Geeignetheit, Notwendigkeit und Verhältnismässigkeit konstruieren. Zur Veranschaulichung wird das Modell auf KI-basierte Fashion-Bildgenerierung bei Deutschlands führendem E-Commerce-Unternehmen OTTO angewendet.

Stichwörter: Künstliche Intelligenz, Verantwortungsvolle KI, KI Ethik, Bildgenerierung, E-Business, Technologiemanagement, Unternehmensethik, Corporate Governance

1 Introduction

Advances in machine learning-based artificial intelligence (AI) create new opportunities and risks for humanity (Coeckelbergh, 2020; Taddeo & Floridi, 2018). For example, the same AI technologies that could improve access to mental healthcare (Zhang & Wang, 2024) have also been reported to encourage violent behavior (Gerken, 2024; Kuznia et al., 2025). This ambivalence raises fundamental normative issues (Coeckelbergh, 2020; Mikalef et al., 2022). Consequently, *responsible* (use of) AI has become a central priority for researchers and practitioners across disciplines (Floridi et al., 2018; Garibay et al., 2023; Ryan & Stahl, 2021).

On the one hand, the existing literature agrees that organizations are crucial to ensuring responsible use of AI (Cihon et al., 2021). On the other hand, there is no consensus on what responsible use of AI actually *is*. Rather, the existing body of knowledge seems to presuppose different conceptualizations of responsibility that share structural similarity with the research on the legitimacy of organizational conduct (Suddaby et al., 2017). One stream of research tends to view responsible use of AI as a *property* that can be specified through principles, requirements, or safeguards and assessed against predefined criteria (Bughin, 2025a, 2025b; Heger et al., 2025; Krijger et al., 2023; Minkkinen et al., 2023). Another stream of research treats the phenomenon as a *process*, suggesting that the responsible use of AI emerges from context-sensitive interpretation and negotiation (Elia et al., 2025; Hagendorff, 2022; Kallina & Singh, 2024; Mittelstadt, 2019; Yilma, 2025).

The issue is that these perspectives rest on conflicting assumptions about the nature of responsibility. Property-oriented views assume that responsibility can be specified independently of situated interpretation, whereas process-oriented views emphasize that responsibility comes into being through such interpretation. Choosing one perspective over the other would simplify the debate, but at the cost of discarding insights that capture important aspects of how responsibility is understood and enacted in organizational practice.

This study addresses this problem by conceptualizing responsible use of AI through the lens of social evaluation as an integrative perspective. Specifically, we combine Bitektine and Haack's (2015) process theory of legitimacy with the principle of proportionality (Karliuk, 2023) to theorize responsible use of AI as both process and outcome of social evaluation across individual and collective levels. We propose a conceptual model ("continuous proportionalization") that captures how organizations form, enact, and revise collective interpretations of responsible use of AI through the convergence of individual propriety judgments along four discursive dimensions: legitimacy, suitability, necessity, and proportionality. For illustration, we apply our model to a case study of fashion image generation at Germany's largest e-commerce company, OTTO.

The remainder of this paper is structured as follows. First, we situate our study in the literature on responsible use of AI and motivate social evaluation as an integrative perspective. We then conceptualize responsible use of AI as continuous proportionalization. Next, we apply the model to the OTTO case for illustrative purposes. Finally, we discuss the contributions, limitations, and implications of this research and offer concluding remarks.

2 Conceptual Background

2.1 Property and Process Perspectives on Responsible Use of AI

Research on responsible use of AI has expanded rapidly in response to the diffusion of AI technologies across organizational contexts (Bach et al., 2025). Private companies, public organizations, and research institutions have issued a growing number of principles, guidelines, and governance frameworks aimed at ensuring ethical or responsible AI use (Jobin et al., 2019). Despite this proliferation of guidance, considerable uncertainty remains about what responsible use of AI means in organizational practice.

As outlined in the introduction, the literature tends to mirror broader debates on the legitimacy of organizational conduct (Suddaby et al., 2017), advancing either property or process perspectives on responsible use of AI.

From a property perspective, responsible use of AI is treated as an attribute of technologies, decisions, or organizational arrangements (Bughin, 2025a, 2025b; Heger et al., 2025; Krijger et al., 2023; Minkkinen et al., 2023). AI use is considered responsible if it satisfies predefined criteria, such as compliance with legal requirements, adherence to ethical principles, or the implementation of technical and organizational safeguards. Responsibility thus appears as something that can be specified ex ante and assessed ex post. This understanding underpins many regulatory and governance approaches that seek to classify AI applications according to their risks and acceptable uses. In this view, responsibility functions as an evaluative label that can be attached to organizational conduct.

In contrast, a process perspective conceptualizes responsible use of AI as something that emerges through ongoing interpretation, deliberation, and justification (Elia et al., 2025; Hagendorff, 2022; Kallina & Singh, 2024; Mittelstadt, 2019; Yilma, 2025). Rather than being a stable property, responsibility is enacted in practice as organizations continuously negotiate what responsible AI use means in light of evolving technologies, shifting regulations, and changing societal expectations. From this perspective, responsibility cannot be fully specified in advance but remains context-sensitive and provisional.

Importantly, these perspectives on the nature of responsible use of AI coexist in tension. Organizations are expected to demonstrate responsibility by pointing to concrete properties (e.g. compliance documentation, safeguards, or compliance structures). Yet, these properties tend to remain incomplete, contested, or insufficient (Mittelstadt, 2019). Especially where established standards lag behind technological development, responsibility cannot be conclusively established through predefined criteria alone (Gogoll et al., 2021). Instead, it must be continuously constructed, justified, and stabilized through processes of deliberation and evaluation (Coeckelbergh, 2024; Gogoll et al., 2021; Watson et al., 2025).

This tension is particularly pronounced in the context of AI, where technological change outpaces the stabilization of normative expectations (Coeckelbergh, 2024; Floridi et al., 2018; Garibay et al., 2023). As a result, responsible use of AI cannot be reduced either to fixed properties or to open-ended processes alone. What remains underexplored is how property and process characteristics are connected in practice. Therefore, we turn to social evaluation as an integrative perspective on the responsible use of AI.

2.2 Social Evaluation as Integrative Perspective

In this paper, we draw on legitimacy theory (Bitektine & Haack, 2015; Suchman, 1995; Suddaby et al., 2017) to integrate the property and process characteristics of responsible use of AI. In legitimacy research, organizational conduct is evaluated not solely based on intrinsic qualities, but from the perspective of relevant audiences who assess whether actions are acceptable, appropriate, and justifiable within a given social context (Suchman, 1995; Suddaby et al., 2017). Legitimacy thus emerges through social evaluation.

Adopting an evaluator perspective (Bitektine & Haack, 2015) allows us to conceptualize responsible use of AI in a way that accommodates both process and property characteristics. Social evaluation foregrounds the processual dimension of responsibility by focusing on how judgments are formed, contested, and revised. At the same time, it explains how the outcomes of these evaluations stabilize responsibility in observable and communicable forms.

It is important to note that legitimacy and responsibility are not equivalent. Organizational conduct may be legitimate without necessarily being responsible in a stronger normative sense. For example, if it complies with regulations or aligns with prevailing norms. Responsibility entails an additional demand: actors must be able and willing to justify their actions by providing reasons that withstand critical scrutiny. The key question is therefore not merely whether AI use is accepted, but which reasons are considered valid in justifying it.

Responsibility thus becomes visible through deliberation, understood as the practice of articulating, weighing, and contesting reasons for action. Crucially, such deliberation is not arbitrary. It is structured by recognizable procedures, standards, and formats that define who evaluates AI use, which considerations are relevant, and how competing reasons are balanced. These structures shape how responsibility is constructed and stabilized in organizational contexts.

Social evaluation provides an integrative lens precisely because it captures this duality. Properties of responsible use of AI can be understood as provisional outcomes of evaluative processes, while these processes themselves are oriented toward producing evaluative outcomes that can guide action. In this sense, responsibility is neither fully given nor endlessly fluid; it is continuously produced, stabilized, and revised through social evaluation.

By adopting an evaluators perspective (Bitektine & Haack, 2015), we can therefore analyze responsible use of AI without reducing it to either fixed normative criteria or open-ended sensemaking processes. Instead, responsible use of AI appears as the outcome of structured social evaluation under conditions of uncertainty, through which organizations seek to render their AI use justifiable to relevant audiences.

In the next section, we build on this perspective to conceptualize responsible use of AI as both process and outcome of social evaluation.

3 Conceptual Model Development

3.1 Core Constructs and Initial Conceptualization

As mentioned above, we derive the core constructs of our conceptualization from the theoretical body on the legitimacy of organizational conduct (Suddaby et al., 2017). We do so because responsible use of AI, like legitimacy, involves a social evaluation of action.

The notion of responsible use of AI presupposes that the decision to (not) use AI is an action that requires justification, and justification requires an audience. Justification entails both the expectation and the obligation to show that an action is taken for good reason. Whether an action is taken for good reason depends on whether it can withstand the scrutiny of those to whom it must be justified. The social nature of action evaluation creates a structural similarity to the legitimacy (Suchman, 1995), which allows us to draw on legitimacy theory (Suddaby et al., 2017) to conceptualize responsible use of AI. While responsibility remains the normative point of reference for our conceptual work, legitimacy theory provides the framework.

We base our work on the multi-level theory of the legitimacy process by Bitektine and Haack (2015). The authors conceptualize the evaluation of organizational conduct as recursive interplay between *individual propriety* and *collective validity* judgments We adopt several other core constructs from Bitektine and Haack (2015) that shape the recursive interplay between individual propriety and collective validity judgement. Which we summarize as follows.

Individual evaluators execute propriety judgments based on two perceptual inputs (Bitektine & Haack, 2015, p. 51). First, they draw on their *perceptions of organizational properties and behavior (Bitektine & Haack, 2015, p. 51)*. Second, they draw on their perceptions of what the collective judgement is (*validity belief*) (Bitektine & Haack, 2015, p. 51). Individual evaluators express their propriety judgement through discourse and *action (Bitektine & Haack, 2015, p. 53)*. As evaluators attend to the behavior of other evaluators (Bitektine & Haack, 2015, p. 51), their actions aggregate into collective effects, and individual propriety judgments converge into validity as a general consensus about the appropriateness of action (Bitektine & Haack, 2015, p. 51; Suddaby et al., 2017, p. 468). This general consensus becomes "institutionalized" (Bitektine & Haack, 2015, p. 53) and reenters the perceptions of individual evaluators as a validity belief, where it will shape and be shaped by future propriety judgment. *Judgment validation institutions* influence this cyclical process by arbitrating among conflicting evaluations and informing the validity beliefs held by individual evaluators (Bitektine & Haack, 2015, pp. 51–52).

Bitektine and Haack (2015) offer a general theory of normative evaluation that is designed to apply across a wide range of evaluative contexts. For the purpose of this research, we introduce several modifications to account for the specificities of responsible use of AI.

First, on the macro level of analysis (Suddaby et al., 2017), we distinguish *organization* and its *environment* as two evaluative spheres. These spheres are connected through observable organizational properties and behavior, which in our case concern the *responsible use of AI*. Individual *perception(s) of organizational AI use* therefore become one of two inputs that *inform individual propriety judgement* and subsequently affect *instance(s) of AI use*.

Second, we introduce *advancement(s) in AI technology* as an additional macro-level construct. This addition reflects that responsible use of AI unfolds within a sociotechnical environment in which normative concerns are continuously reshaped by rapid technological development (Mikalef et al., 2022; Stahl, 2012).

These specifications yield an initial conceptualization of responsible use of AI as process and result of social evaluation (Figure 1).

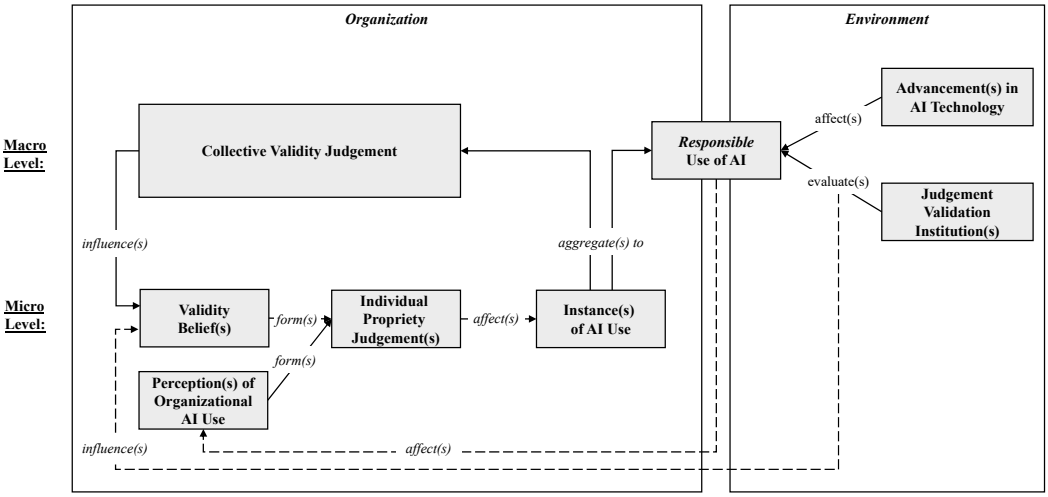


Figure 1: Initial Conceptualization of Responsible Use of AI as Social Evaluation

While this initial conceptualization clarifies how individual and collective evaluations interact and change over time, it leaves open an essential question: how do individual propriety judgments converge into a collective validity judgement? Bitektine and Haack (2015) provide the important insight that this aggregation happens discursively. Building on this, we adopt the principle of proportionality (Karliuk, 2023) to identify discursive dimensions through which individual propriety judgements translate into collective validity judgment of responsible use of AI.

3.2 Leveraging the Principle of Proportionality

The principle of proportionality is a long-standing heuristic for evaluating the validity of action. It originated in Prussian administrative law to assess the legitimacy of government intervention in economic and social affairs (Cohen-Eliya & Porat, 2010). It was based on the emerging understanding that citizens are legal entities with individual rights, and that state actions restricting these rights can only be carried out if (and to the extent to which) they are justified (Cohen-Eliya & Porat, 2010). To this end, Prussian court judges tested state action for *Legitimacy*, *Suitability*, *Necessity*, and *Proportionality* in the strict sense. Progressing through this sequence of tests (Figure 2) enabled them to determine whether state action is justified in the sense that its end(s) are legitimate and its mean(s) adequate (Cohen-Eliya & Porat, 2010; Sobek & Montag, 2018).

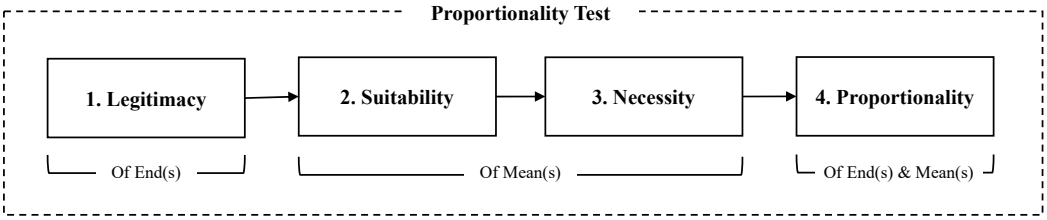


Figure 2: Proportionality Test Components

Today, proportionality is a cornerstone of constitutional and human rights law across the globe (Alexy, 2014). This expansion was possible, because the principle of proportionality and the proportionality test are sufficiently abstract. Conceptually, there is no reason not to apply it beyond the legal domain to evaluate actions that affect others.

In this spirit, Karliuk (2023) suggested reconceptualizing proportionality for AI ethics. Specifically, for decisions regarding "both to using AI as such and, if a decision is made to use it, to choose the right AI method" (Karliuk, 2023, p. 989). Applying the proportionality principle for the responsible use of AI means "addressing AI systems in a way that […] their use do[es] not exceed what is necessary to achieve legitimate aims" (Karliuk, 2023, p. 988). It follows to consider the use of AI responsible if (and only to the extent to which) it is *suitable*, *necessary*, and *proportionate* to achieve a *legitimate* aim.

We agree that the proportionality test offers valuable "structural guidance […] to reach […] justifiable decision[s]" and could therefore "play an important role in AI ethics" (Karliuk, 2023, p. 987). Its usefulness for our work lies in the fact that it makes the basic structure of practical reasoning explicit. When people scrutinize the legitimacy of action, they intuitively ask questions like: Was it done for the right reasons? Was this the appropriate way to pursue the goal? Were the means and ends in balance? The proportionality principle captures these intuitions in a systematic way.

If we assume that the organizational use of AI is, like any other action, subject to justification, then these same questions naturally arise in deliberation about AI. The proportionality principle therefore provides a set of shared evaluative dimensions that are likely to surface in discursive exchanges, and through which evaluators compare and scrutinize their propriety judgments against those of others. If Bitektine and Haack (2015) are correct that collective validity judgement emerges discursively, then this convergence will plausibly be based on shared understandings of legitimacy, suitability, necessity, and proportionality.

In this sense, the proportionality principle offers a plausible account of how individual propriety judgments may converge into a collective validity judgement of responsible use of AI. We refer to this discursive convergence as *proportionalization*.

Proportionalization fills the conceptual gap in our initial conceptualization of responsible use of AI as social evaluation. We present our extended conceptual model in the next section.

3.3 Responsible Use of AI as Continuous Proportionalization

We conceptualize responsible use of AI in organizations as a process and result of recursive interplay between individual propriety and collective validity judgment in a dynamic, socio-technical environment (Figure 3).

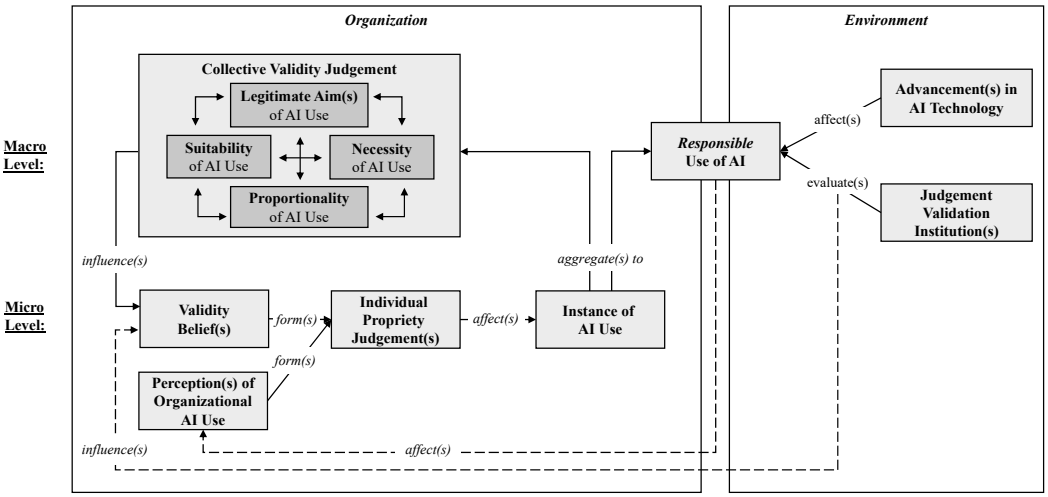


Figure 3: Responsible Use of AI as Continuous Proportionalization

In this understanding, organizational members act as evaluators of the responsible use of AI. Based on their perception(s) of organizational AI use and their validity belief(s) about the prevailing collective judgement, the members of an organization execute propriety judgement(s), assess the normative acceptability of AI use, and influence how AI is (not) used in a given context. Decisions and actions at the individual aggregate at the organizational level. As evaluators observe how others act and justify their decisions, individual propriety judgement(s) concerning responsible use of AI converge into collective validity judgement(s). That is, a shared understanding at the organizational level of when the use of AI is legitimate, suitable, necessary, and proportionate. Once emerged, this collective understanding becomes institutionalized and reenters the practical consciousness of evaluators as validity belief(s), where it will shape and be shaped by future individual propriety judgement(s).

This recursive interplay unfolds within the organization but is influenced by its dynamic sociotechnical environment. Judgement validation institution(s) (e.g., regulators, courts, standards bodies, NGOs, and media) evaluate the use of AI and provide cues about when they consider AI use legitimate, suitable, necessary, or proportionate. These validity cues influence the validity belief(s) held by individual evaluators and can indirectly shape both the collective validity judgement and the organizational use of AI. The same is true for technological advancement(s), which open new possibilities for using (new) AI technologies that raise (new) normative questions.

We therefore suggest viewing responsible use of AI as both the process and the result of *continuous proportionalization*. In the next section, we apply continuous proportionalization as an analytical lens in an illustrative case study.

4 Empirical Illustration

The primary contribution of this paper is conceptual. The case evidence is used to demonstrate the empirical plausibility of continuous proportionalization and to illustrate how the evaluative mechanisms specified by our model (i.e., recursive interplay of individual

propriety and collective validity judgment) can be traced in organizational practice. While capturing the full contextual range of continuous proportionalization would require longitudinal research, the present case allows us to observe key dynamics and evaluative dimensions in situ.

4.1 Case Study Research Method

In line with the illustrative function, we conducted a propositional single case study (Yin, 1994). We chose to study the use of AI at Germany's largest e-commerce company OTTO. We selected OTTO because its retail platform business relies heavily on AI (Christophersen & Pärn, 2021), and the company positions itself as a responsible AI user. As the CIO explained: *"We are very much determined by the values of our owner family. They have given us a Code of Ethics on how we want to act as a company [...]. One sub-category […] is the responsible use of technology. [...] We are already making extensive use of artificial intelligence and [...] are giving this a lot of thought"(Otto, 2021)*.

We decided to focus our case study on OTTO's AI use for fashion image generation. This is because fashion image generation promises significant business value for e-commerce companies like OTTO. However, fashion articles are typically displayed on human bodies, which raises normative concerns and subjects fashion image generation to a certain level of public scrutiny. As evidenced by the recent controversy regarding Guess's decision to use AI models in an international fashion campaign (Rufo, 2025). In addition, fashion image generation raises general concerns of intellectual property, bias, and environmental impact (Katirai et al., 2024). Hence, OTTO faces a difficult question: how to use AI in fashion image production responsibly?

Data Type	Σ	Duration
Key Informant Interviews	**4**	**196 min**
(ID-1) Digital Content Management Expert	1	49 min
(ID-2) Teamlead Content Operations	1	47 min
(ID-3) Digital Content Production Specialist	1	46 min
(ID-4) Product Owner CGI Production Service	1	54 min
Participant Observation	**3**	**270 min**
Internal AI Use Case Assessment Workshops	3	270 min
Documents & Archival Records	**13**	**39 min**
Internal Company Documents & Intranet Websites	12	-
Podcasts	1	39 min

Table 1. Case Data by Data Type

We collected data on OTTO's use of AI in fashion image generation from interviews, participant observation, documents & archival records (Table 1). We purposefully selected interviewees who spearheaded or managed AI initiatives in OTTO's image content production because they represented different functions along the fashion image production

and AI-use-case evaluation process (content operations, production, product ownership, and digital content management). All interviews were conducted via Microsoft Teams, recorded, and transcribed. They lasted between 46 and 54 minutes and were conducted in German. We used the speech recognition system "Whisper" for transcription (Radford et al., 2022) and corrected the output as needed. We also collected data by observing three workshops in which OTTO assessed potential AI use cases for image generation. In these workshops, we acted as non-intervening participant observers and did not influence discussions or decision-making processes. Observing the workshops allowed us to capture firsthand how discursive interactions between individual propriety judgments and collective validity judgments unfold. In particular, how participants articulated, contested, and aligned reasoning. We triangulated interview accounts and workshop observations with information from internal documents (e.g., PowerPoint presentations, online whiteboards) and archival records (e.g., intranet pages, podcasts).

For data analysis, we stored the case evidence from all sources in an online database (ATLAS.ti). We used qualitative methods and analyzed the data in four steps. First, we marked all case data that related to fashion image generation and responsible use of AI. Which resulted in over 120 quotes. Then, we coded these quotes deductively using the main constructs of our conceptual model following Yin (1981, 1994). This deductive coding allowed us to group case evidence from different sources under common categories (e.g. individual propriety judgement[s]). Next, we applied inductive coding to abstract and capture additional insights from the quotes. Finally, we combined the coding results and interpreted them using our case knowledge. Table 2 exemplifies how we marked, grouped, coded, and interpreted our data.

Data Source	Data Excerpt	Deductive Code(s)	Inductive Code(s)	Interpretation
Interview (ID-1)	*"And that's exactly what it is: subjective perception. When you start* ***discussing*** *something like this with a lot of people,* ***everyone brings their own subjective view*** *to the table. And* ***stepping outside*** *your own perception and trying to* ***discuss things objectively is the biggest challenge*** *for everyone."*	Individual Propriety Judgement(s)	Subjective Perception Plurality of Perspective Difficulty: Discursive Objectivation	The codes indicate that the democratization of AI image generation increases the number of *individual propriety judgment(s)*. This plurality of subjective perspectives seems to generate normative uncertainty and complicate the discursive convergence.
Interview (ID-3)	*"[...] it was a bit like the* ***Wild West because content production is becoming totally democratized****. I'm a trained expert, I studied it and worked my way through the whole subject area. And now there's someone who comes from the finance sector, but has a good imagination and knows how to write a prompt – they can produce just as good content."*	Individual Propriety Judgement(s)	Democratization of Image Production Normative Uncertainty ("Wild West")	

Data Source	Data Excerpt	Deductive Code(s)	Inductive Code(s)	Interpretation
Interview (ID-1)	*"The inspiration comes from both sides. You could say: Okay, there's the* ***competition****. OTTO wants to be part of the competition. The* ***competition is pretty fast when we measure ourselves against it.****"*	Perception(s) of Organizational AI Use	Monitoring of Competition Benchmarking	The codes indicate that *perception(s) of organizational AI use* are shaped by constant monitoring of the organizational environment. In addition to technological advancements, close observation of competitors and benchmarking against their practices play a central role.
Interview (ID-4)	*"A lot of it is driven internally. We have highly* ***motivated colleagues*** *who are* ***constantly monitoring what is happening*** *out there, what* ***new technologies*** *are available, what we can do, and what the* ***competition*** *is doing."*	Perception(s) of Organizational AI Use	Continuous Monitoring Monitoring of Competition Monitoring of Technology	
Interview (ID-1)	*"But* ***our utility analysis*** *[...] really helps us prioritize: What do we tackle next, and what don't we tackle? [...] Otherwise, you don't know how to decide. At least, I wouldn't know."*	Proportionality of AI Use	Institutionalization of Proportionality Utility Analysis as Decision Instrument	The codes indicate that a shared understanding of the *proportionality of AI use* tends to be institutionalized through formal decision instruments, such as utility analyses, which help decision makers compare AI use cases and impose order under conditions of limited organizational resources. At the same time, institutionalized criteria remain open to revision and may change over time.
Interview (ID-4)	*"There are lots of great things you could do. And, of course,* ***everyone has their own baby or pet project.*** *To help us bring a little* ***order*** *to the process, we also have* ***our utility analysis****."*	Proportionality of AI Use	Utility Analysis as Decision Instrument Ordering of AI Use Cases	
Participant Observation [Workshop]	*During a workshop, we observed practitioners applying OTTO's utility analysis to evaluate and select AI use cases. However, the predefined scale(s) for assessing business value did not allow for meaningful differentiation between the potential AI use cases in fashion image generation. The* ***participants discussed and adapted the scale*** *to better reflect the specific business context.*	Proportionality of AI Use	Institutionalization of Proportionality Utility Analysis as Decision Instrument Adaptation of Evaluation Scheme	

Table 2. Example of Case Data Analysis

We made sense of the data until the research team agreed that explaining the use of AI in fashion image generation as a process and result of continuous proportionalization is free from obvious absurdities and thus sufficiently plausible (Polanyi, 1962). Finally, we discussed our findings with OTTO practitioners for external validation.

Having described the research method and introduced the case company, we now provide additional background to understand OTTO's use of AI in fashion image production.

4.2 How OTTO Uses and Produces Image Content

Digital images are essential to e-commerce retailers like OTTO, as they primarily engage with customers through electronic channels. Unlike in brick-and-mortar retail, where customers can usually look at or touch products beforehand. As an interviewee explained: "*As humans, we are visual beings. We need information about a product to be presented*

not only in text form, but also visually, for us to be able to imagine it. […] Hence, image content is particularly important in facilitating purchasing decisions." (ID-3)

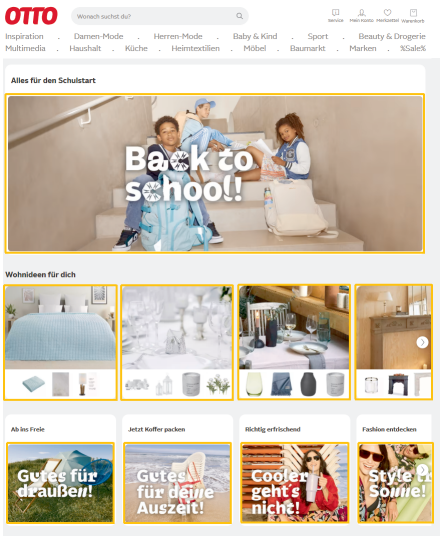
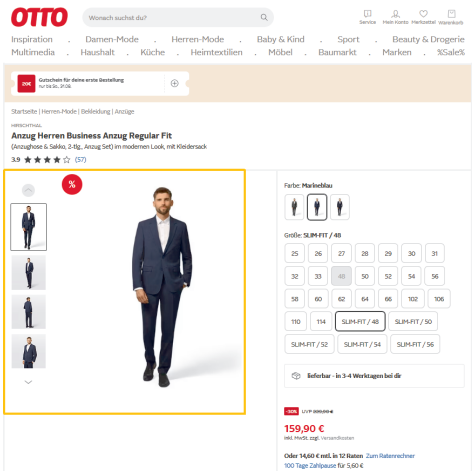
Type	Marketing Content	Product Content
Function	Awareness & Inspiration	Activation & Decision
Use (e.g.)	Landing Pages, Campaign Hubs, Inspiration Widgets, Editorial Content	Product Detail Pages (incl. Variations, Bundles, Configurators)
Image Example		

Figure 4: Image Content Types

OTTO differentiates between *marketing* and *product* images (Figure 4). Marketing images are used in the early phases of the customer journey to position the OTTO brand, create desire, and inspire customers. Product images are used in later stages to help customers find specific items, gain orientation, and compare alternatives.

The main difference between marketing and product images is the degree of exactness with which they must represent the products sold on OTTO's website ("article fidelity"). Unlike marketing, product images require a high level of article fidelity. After all, product images are supposed to help customers imagine the product without creating unrealistic expectations.

Until recently, OTTO used either photography or computer-generated imagery (CGI) to produce digital images. Generative AI models such as Midjourney and Stable Diffusion added a third possibility. They can create digital images from textual inputs by predicting likely pixel sequences (Saharia et al., 2022). Unlike photography and CGI, image generation models do not require real or digital objects. Which promises new opportunities to produce image content faster, cheaper, and more flexible. In the next section, we apply our conceptual model to illustrate the responsible use of AI in OTTO's fashion image production.

5 Responsible Use of AI for Fashion Image Production at OTTO

For OTTO, the question of responsible use of AI in fashion image production emerged with the rise of image generation models in 2023. Since then, OTTO progressed through what can be described as two *proportionalization cycles*. Below, we outline these cycles and provide examples of AI use cases that OTTO put into production based on its proportionalized understanding of responsible use of AI.

5.1 First Proportionalization Cycle

5.1.1 Formation of Individual Propriety Judgement(s)

At OTTO, individual propriety judgements initially formed through observation and perception of the socio-technical environment in which the organization operates. These perceptions were shaped not only by rapid technological developments in generative AI, but also by close monitoring of competitors' activities. As an interviewee noted: "*We have highly motivated colleagues who are constantly monitoring what is happening out there, what new technologies are available, what we can do, and what the competition is doing*." (ID-4)

Based on these emerging propriety judgements, AI image generation developed organically within OTTO, giving rise to multiple initiatives across the organization. The low entry barriers to the technology enabled many employees to experiment with AI image generation, including individuals outside the units traditionally responsible for content production: *"[...] it was a bit like the Wild West because content production is becoming totally democratized. I'm a trained expert, I studied it and worked my way through the whole subject area. And now there's someone who comes from the finance sector, but has a good imagination and knows how to write a prompt – they can produce just as good content." (ID-3)*

5.1.2 Formation of Collective Validity Judgement(s)

As increasing numbers of employees at OTTO engaged with AI image generation and observed how others expressed propriety judgements, normative questions emerged regarding how the technology should be used. These questions triggered extensive discursive exchanges across hierarchical levels: *"Discussions that were held naturally included questions like, 'How perfect do our people actually look?' I still remember that when we were with the executive board, this question was raised as well. […] From my perspective, this is a secondary issue, because even today we do not photograph the average person next door, but professional models who go through casting processes beforehand." (ID-3)*

Through these discursive exchanges, individual propriety judgements began to converge into collective validity judgements regarding the responsible (use of AI) in fashion image generation. Interviewees emphasized, however, that this convergence was neither immediate nor straightforward. The large number of involved actors and the diversity of perspectives made it difficult to establish common understanding. Nevertheless, over time, discussions increasingly crystallized around questions of the legitimacy, suitability, necessity, and proportionality.

(1) Legitimacy of AI Use: The purpose of AI use in fashion image production is to obtain "*inspiring content for customers cheaply and quickly*" (ID-1). OTTO assessed the

legitimacy of this purpose with respect to internal expectations and regulatory validity clues. As for regulatory validity clues, the EU AI Act classifies fashion image generation as a "limited risk" application, which entails transparency and labelling requirements but does not constitute a prohibited use of AI (European Commission, 2021). Internally, fashion image generation aligns with OTTO's general business interests. Therefore, fashion image generation was considered legitimate, provided it complies with OTTO's guidelines for image production, which would prohibit, for example, the generation of visuals that are offensive or discriminatory.

(2) Suitability of AI Use: OTTO determined the suitability of AI use for fashion image production primarily through experimentation: "*We explored different AI services and asked ourselves, what happens when we input images? […]. What kind of results do we get? […] And so, step by step, we keep sorting things out: This works, that doesn't. This works, that doesn't…"* (ID-1).

This experimentation took place in the context of rapid technological development: "*AI […] gives us new possibilities every week […]. The pace is enormous, and I have rarely seen an exponential growth curve as clearly.*" (ID-2). Early experiments revealed limitations such as anatomical inaccuracies in depictions of human figures, which were gradually reduced as the technology improved. At the same time, OTTO encountered situations in which technological possibilities shifted in unexpected ways: "*We were surprised to find that the quality of lingerie images was better than that of ordinary textiles. […] However, it is almost impossible to produce lingerie images with current models from Google and others. Because they implemented filter functions […]. So, things looked very promising regarding lingerie images for a while, but possibilities are currently very limited again." (ID-2)*

Through such experimentation, OTTO developed a general understanding of the types of problems for which AI is particularly suitable. As one interviewee summarized: "*Generative AI excels at predicting […] which pixel is most likely to appear in a particular spot. However, when it comes to representing articles accurately, […] you must 'bend' AI models for this purpose. Because they are not built to reproduce something exactly […]. Which is why we still face clear limitations when it comes to article fidelity." (ID-2)*. OTTO therefore concluded that current image generation models are more suitable for producing content with low article fidelity (i.e., marketing images) and less suitable for content that requires high article fidelity (i.e., product images).

(3) Necessity of AI Use: OTTO assessed the necessity of AI use by comparing it with alternative means of image production such as photography and CGI. In this comparison, AI was perceived as "*almost unbeatable in terms of efficiency*" when creating marketing content (ID-2). At the same time, AI was not seen as merely replacing existing processes but also as enabling new applications that had previously been infeasible. As one interviewee explained: "*There are uses that would not have been possible […] before. There simply would have been no image. That would have been the solution.*" (ID-2). Accordingly, OTTO developed the understanding that AI may be necessary in "*problem cases*" (ID-4) where photography or CGI fall short.

(4) Proportionality of AI Use: At OTTO, the use of AI needed to be justified in terms of the company's top priority to "*do the things that take us furthest forward*" (ID-4). To operationalize this maxim, OTTO developed a "*utility analysis*" (ID-1) that enables deci-

sion-makers to weigh all relevant considerations for a specific AI use case and compare it to others.

The utility analysis consists of more than 20 criteria (e.g., business benefits, strategic alignment, and risks) that OTTO identified as relevant for evaluating AI use cases. These criteria were assigned individual weights by systematically comparing their relative importance. Based on the resulting scores, use cases were classified within a portfolio and compared against one another to support prioritization decisions.

The potential AI use cases for fashion image generation were also supposed to be assessed using this instrument. Because "*There are lots of great things you could do. And, of course, everyone has their own baby or favorite project. We have the utility analysis to bring a little order to things*" (ID-4). During one workshop, however, we observed that the predefined scales for assessing business value did not allow for meaningful differentiation among fashion image generation use cases. In response, participants discussed and adapted the evaluation scale to better reflect the specific business context.

Based on the utility analysis, OTTO put several AI use cases for fashion image generation into production. Three of which we describe below.

5.1.3 Examples of AI Use in OTTO's Fashion Image Production

The first AI use case that OTTO deemed justifiable concerned the generation of marketing images. Because article fidelity plays only a secondary role in this context, AI was assessed as suitable for producing visuals whose primary purpose is to inspire customers, encourage engagement, and increase the likelihood of sales. Traditional production processes, such as model and location photography, were highly demanding in terms of effort, cost, and scalability. The use of AI was therefore considered both necessary and proportionate, as it enabled the efficient creation of inspiring content at approximately 60 percent lower cost than conventional image production. Based on this assessment, OTTO established a fully digital workflow in which marketing images are generated entirely by AI.

Figure 5: AI-Generated Marketing Image on OTTO Landing Page

In contrast to marketing visuals, product images require high article fidelity. OTTO therefore uses AI as a complementary tool alongside established production methods, rather than as a full replacement (Figure 6).

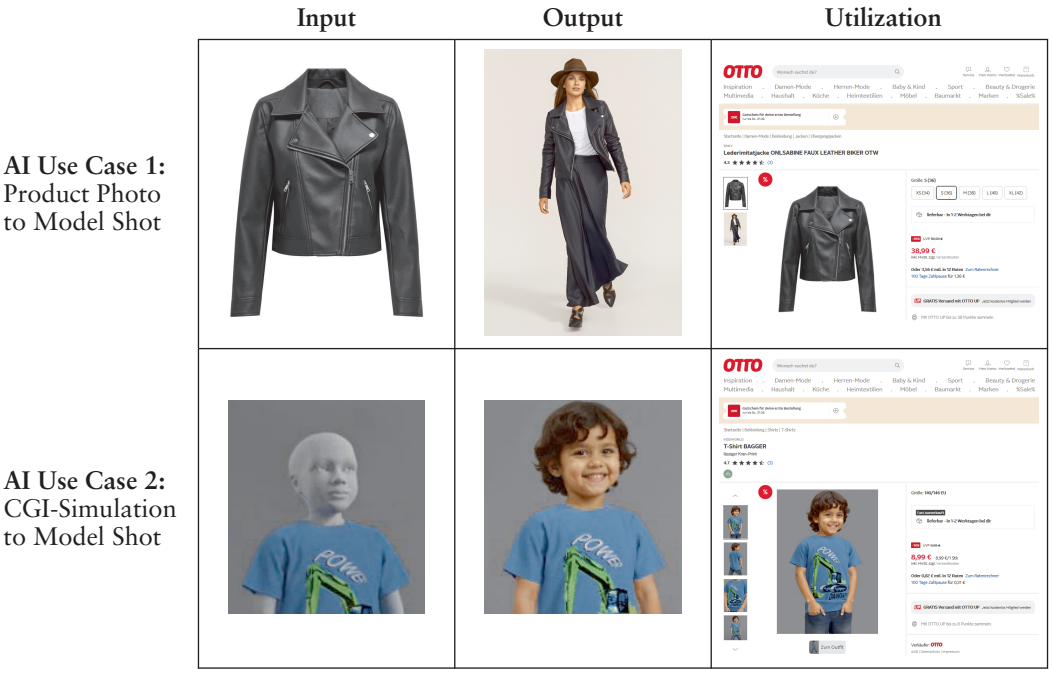


Figure 6: AI-Enhanced Product Images

One such use case involves generating model images based on existing product photos. AI is applied where model shots are unavailable due to logistical or economic constraints. As one interviewee noted, traditional model photography requires an "*insanely high level of effort*" (ID-1) in terms of sample logistics, studio costs, and personnel coordination. In this context, OTTO found AI suitable for generating front-view model images from product photos, while manual post-production ensures that no counterfactual or misleading images are created.

A related use case builds on CGI simulations from OTTO's fashion product development process. While CGI enables "*maximum product fidelity*" (ID-3), it has "*limitations when it comes to depicting people*" (ID-3). Here, AI allows OTTO to generate realistic model images for fashion articles that have not yet been physically produced. This combination enables earlier visualization of products while maintaining control over article fidelity.

5.1.4 Institutionalization of Collective Validity Judgement(s)

Once established, the collective understanding of responsible use of AI in fashion image generation became institutionalized at OTTO. One prominent form of institutionalization is the expansion of the utility analysis as an agreed decision instrument across multiple initiatives. As one interviewee explained: "*We now use the utility analyses across different initiatives. But we have also noticed that […] not all teams in the company are using it yet.*" (ID-1)

Beyond formal decision instruments, we observed additional forms through which the collective understanding of responsible AI use becomes institutionalized in everyday practice. These include the development of standardized prompts and curated pools of AI models for image generation. As one interviewee noted: "*For example, what do future prompts need to look like so that we really only get these kinds of models? […] Is there also an AI model pool that people can draw on? […] These are the steps we are thinking about now after this MVP phase… How can we bring more stability into this and more control over what is generated?"* (ID-1)

At the same time, ongoing changes in the socio-technical environment have prompted OTTO to enter a subsequent proportionalization cycle that is still unfolding.

5.2 Second Proportionalization Cycle

5.2.1 Expansion of the Legitimacy Object(s)

The second proportionalization cycle emerged from organizational learning about the technical possibilities of AI image generation in the first cycle. Compared to the initial phase, the legitimacy object shifted in a fundamental way. While AI had initially been discussed primarily as an alternative means for established purposes, such as replacing traditional model photography, OTTO began to consider whether AI could also be used for new purposes that had not previously been part of image production practices.

One example concerns the exploration of AI image generation for baby fashion. After successfully generating model images for adult fashion items based on product photos, OTTO began examining whether a similar approach could be applied to baby clothing. Because babies had not been photographed before, this exploration marked an expansion of the legitimacy object and opened a qualitatively new field of application.

In contrast to adult fashion, however, generating synthetic images of babies raised not only technical challenges related to article fidelity but also heightened ethical and normative concerns. A central issue concerned compliance with internal guidelines on nudity, which stipulate that "*too much skin should not be shown*" (ID-1). At the same time, the absence of established production practices for baby images limited the availability of reference points. Additional concerns emerged regarding the potential misuse of synthetic baby images. As one interviewee explained, no collective consensus has yet emerged regarding responsible use of AI in this context: "*It is definitely a case. There is a need and we are moving forward. But only within the guidelines. And this is where the ethical aspect comes in: not everything at all costs, not just focusing on the business case […] but even if they are synthetic people, portraying them with dignity. [...] We cannot yet say what this will look like in the end, because we are only just starting.*" (ID-1)

5.2.2 Socio-Technical Dynamics in OTTO's Environment

The ongoing proportionalization cycle is further shaped by judgement validation institution(s) in OTTO's environment. In addition to a dynamic regulatory landscape, anticipated customer reactions are particularly salient for how OTTO reflects on what it may hold to be a legitimate, suitable, necessary, and proportionate use of AI over time.

Under current legal conditions, OTTO does not yet visibly mark AI-generated images as such. This will change as the EU AI Act enters into force. At present, however, "*for the*

customer, it is not distinguishable whether an article was photographed on a real model or generated by AI" (ID-3).

In line with our main proposition, the collective understanding of responsibility at OTTO that has emerged through proportionalization and currently informs the use of AI in fashion image generation remains provisional. It guides organizational action in the present, yet remains open to revision as new evaluative inputs emerge from the socio-technical environment. As one interviewee concluded: "*And this is now the point where we say: We generate the [image] content, try to design it as well as possible in the sense of OTTO. And once the content is on the website, we go into surveying […] to include our customers' perceptions in further development. […] It could be that everyone says: 'Oh God, we don't want this at all.' Then we would have to stop the project and say: we no longer generate images, we go back to photography. […] But to even get into this try-and-error, the images have to go live*" (ID-1).

6 Discussion and Conclusion

We conducted this research to advance our understanding of responsible use of AI in organizations. To this end, we integrated Bitektine and Haack's (2015) process theory of social evaluation with the principle of proportionality (Karliuk, 2023). This integration resulted in a conceptual model ("continuous proportionalization") that captures how organizations form, enact, and revise collective interpretations of responsible use of AI in a dynamic sociotechnical environment. For illustration, we applied the model to a case study of AI use in fashion image generation at Germany's largest e-commerce company, OTTO.

With this research, we make three interrelated contributions to the growing body of literature on governance and responsible use of AI in organizations.

First, we offer an integrative perspective on the nature of responsible use of AI. Prior research has tended to conceptualize responsible use of AI either as a relatively stable property of systems, principles, or organizational arrangements (Bughin, 2025a, 2025b; Heger et al., 2025; Krijger et al., 2023; Minkkinen et al., 2023), or as an ongoing process of deliberation, reflection, and adaptation (Elia et al., 2025; Hagendorff, 2022; Kallina & Singh, 2024; Mittelstadt, 2019; Yilma, 2025). This has left a theoretical puzzle as to how responsibility can be both dynamic and, at least temporarily, stabilized in organizational practice. By conceptualizing responsible use of AI as a process of social evaluation (Bitektine & Haack, 2015), we provide an integrative stance that accounts for both its processual and property-like characteristics without collapsing into either.

Second, we theorize how shared organizational understandings of responsible use of AI emerge. Prior work has highlighted the importance of discourse and deliberation in responsible AI (Coeckelbergh, 2024; Gogoll et al., 2021; Watson et al., 2025). We extend this literature by introducing a process model that explains how individual propriety judgments converge discursively into collective validity judgments within a dynamic sociotechnical environment. Our model further shows that this convergence is not arbitrary, but crystallizes around interpretations of legitimacy, suitability, necessity, and proportionality of AI use. In doing so, we elevate the principle of proportionality from a situational decision test (Karliuk, 2023) to a set of shared discursive dimensions through which collective interpretations of responsible use of AI are formed, stabilized, and revised over time.

Third, we illustrate the value of drawing on legitimacy and institutional theory for research on responsible use of AI. As other scholars have noted (de-Lima-Santos et al., 2025; Horneber, 2025), legitimacy and institutional theory provide useful lenses for conceptualizing responsible use of AI. While beyond the scope of this paper, studying dynamic sociotechnical phenomena such as responsible use of AI may offer opportunities not only to draw on, but also to contribute back to legitimacy and institutional theory.

We argue that our findings are transferable beyond the empirical context examined in this study (i.e., e-commerce). This is because our conceptual model is grounded in general theory of the legitimacy of organizational conduct (Bitektine & Haack, 2015) and a broadly applicable heuristic of action validation (Karliuk, 2023). It follows that challenging the explanatory power of continuous proportionalization in other settings offers promising avenues for future research. On the one hand, the model could be applied across different industries and regional contexts to examine how proportionalization is enabled or constrained under varying regulatory, competitive, or cultural conditions. On the other hand, the model could be extended to high-stakes domains of AI use, such as healthcare or military applications. While we expect the recursive interplay between individual propriety and collective validity judgments to persist, the dynamics of proportionalization may differ as the severity of potential consequences increases.

Additional research opportunities arise from the limitations of our study. While our case study illustrates the recursive interplay between individual propriety judgments and collective validity judgments, it does not capture proportionalization over an extended period or in all its contextual nuances. Future research could therefore adopt a longitudinal perspective, to trace how proportionalization unfolds over time (e.g. using ethnographic methods). Moreover, our study did not assess the effects of proportionalization on measurable outcomes (e.g., regulatory compliance and organizational performance). Future research could develop and test metrics that capture both the effectiveness and the ethical implications of responsible AI use. Finally, although we identify practical challenges in the discursive convergence of propriety judgments, we did not examine how such challenges might be addressed. This opens avenues for design-oriented research to develop tools, processes, or governance mechanisms that support organizations in forming shared understandings of responsible use of AI.

Our research offers actionable implications for managers. We suggest that responsible use of AI is neither a one-time compliance exercise nor something that can be "solved" by issuing principles or guidelines alone. Instead, it requires continuous attention and the active organization of discourse through which collective interpretations of responsible use of AI will be formed, challenged, and revised. Creating and sustaining such discursive spaces then becomes a central managerial task. At the same time, our findings imply that responsible use of AI cannot simply be imported from external standards or best practices. While external cues such as regulation or industry norms matter, a shared understanding of responsibility must be worked out internally, in relation to an organization's specific goals, constraints, and technologies. To support this process, our conceptual model provides practitioners with a set of orienting questions around which responsible use of AI crystallizes: Is the aim for using AI legitimate? Is AI suitable for achieving it? Is the use of AI necessary? And are means and ends proportionate? In discussions with OTTO practitioners, these dimensions were perceived as intuitive and useful for structuring delib-

eration. They may help managers navigate the difficult task of creating conditions under which continuous proportionalization can unfold.

In conclusion, this research contributes to a more nuanced understanding of what responsible use of AI is and how it can be organized. Which we trust other researchers and practitioners may find useful in their shared efforts to ensure that AI technologies are used for good.

References

Bach, T. A., Kaarstad, M., Solberg, E., & Babic, A. (2025). Insights into suggested Responsible AI (RAI) practices in real-world settings: a systematic literature review. *AI and Ethics*, *5*(3), 3185–3232. https://doi.org/10.1007/s43681-024-00648-7

Bitektine, A., & Haack, P. (2015). The "macro" and the "micro" of legitimacy: Toward a multilevel theory of the legitimacy process. *Academy of Management Review*, *40*(1), 49–75. https://doi.org/10.5465/amr.2013.0318

Bughin, J. (2025a). Doing versus saying: responsible AI among large firms. *AI & Society*, *40*(4), 2751–2763. https://doi.org/10.1007/s00146-024-02014-x

Bughin, J. (2025b). The role of AI assets and capabilities in shaping responsible AI deepening: a random forest machine learning view. *AI and Ethics*, *5*(6), 6313–6327. https://doi.org/10.1007/s43681-025-00802-9

Christophersen, T., & Pärn, J. (2021). Data Science bei OTTO. In P. Buxmann & H. Schmidt (Eds.), *Künstliche Intelligenz* (pp. 101–115). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61794-6_6

Cihon, P., Schuett, J., & Baum, S. D. (2021). Corporate governance of artificial intelligence in the public interest. *Information (Basel)*, *12*(7), 275. https://doi.org/10.3390/info12070275

Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.

Coeckelbergh, M. (2024). Artificial intelligence, the common good, and the democratic deficit in AI governance. *AI and Ethics*, *5*(2), 1491–1497. https://doi.org/10.1007/s43681-024-00492-9

Cohen-Eliya, M., & Porat, I. (2010). American balancing and German proportionality: The historical origins. *International Journal of Constitutional Law*, *8*(2), 263–286. https://doi.org/10.1093/icon/moq004

de-Lima-Santos, M.-F., Yeung, W. N., & Dodds, T. (2025). Guiding the way: a comprehensive examination of AI guidelines in global media. *AI & Society*, *40*(4), 2585–2603. https://doi.org/10.1007/s00146-024-01973-5

Elia, M., Ziethmann, P., Krumme, J., Schlögl-Flierl, K., & Bauer, B. (2025). Responsible AI, ethics, and the AI lifecycle: how to consider the human influence? *AI and Ethics*, *5*(4), 4011–4028. https://doi.org/10.1007/s43681-025-00666-z

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Garibay, O. O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S. M., Garibay, I., Grieman, K., Havens, J. C., Jirotka, M., Kacorri, H., Karwowski, W., Kider, J., Konstan, J., Koon, S., Lopez-Gonzalez, M., Maifeld-Carucci, I., … Xu, W. (2023). Six Human-Centered Artificial Intelligence Grand Challenges. *International Journal of*

Human–Computer Interaction, *39*(3), 391–437. https://doi.org/10.1080/10447318.2022.2153320

Gerken, T. (2024, December 11). Chatbot "encouraged teen to kill parents over screen time limit." *BBC News*. https://www.bbc.com/news/articles/cd605e48q1vo

Gogoll, J., Zuber, N., Kacianka, S., Greger, T., Pretschner, A., & Nida-Rümelin, J. (2021). Ethics in the Software Development Process: from Codes of Conduct to Ethical Deliberation. *Philosophy & Technology*, *34*(4), 1085–1108. https://doi.org/10.1007/s13347-021-00451-w

Hagendorff, T. (2022). A virtue-based framework to support putting AI ethics into practice. *Philosophy & Technology*, *35*(3), 1–24. https://doi.org/10.1007/s13347-022-00553-z

Heger, A. K., Passi, S., Dhanorkar, S., Kahn, Z., Wang, R., & Vorvoreanu, M. (2025). Towards a Responsible AI Organizational Maturity model. *Proceedings of the ACM on Human-Computer Interaction*, *9*(7), 1–33. https://doi.org/10.1145/3757514

Horneber, D. (2025). Understanding the implementation of responsible artificial intelligence in organizations: A Neo-institutional theory perspective. *Communications of the Association for Information Systems*, *57*, 8. https://doi.org/10.17705/1cais.05708

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kallina, E., & Singh, J. (2024). Stakeholder involvement for responsible AI development: A process framework. *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–14. https://doi.org/10.1145/3689904.3694698

Karliuk, M. (2023). Proportionality principle for the ethics of artificial intelligence. *AI and Ethics*, *3*(3), 985–990. https://doi.org/10.1007/s43681-022-00220-1

Katirai, A., Garcia, N., Ide, K., Nakashima, Y., & Kishimoto, A. (2024). Situating the social issues of image generation models in the model life cycle: a sociotechnical approach. *AI and Ethics*, *5*(2), 1769–1786. https://doi.org/10.1007/s43681-024-00517-3

Krijger, J., Thuis, T., de Ruiter, M., Ligthart, E., & Broekman, I. (2023). The AI ethics maturity model: a holistic approach to advancing ethical data science in organizations. *AI and Ethics*, *3*(2), 355–367. https://doi.org/10.1007/s43681-022-00228-7

Kuznia, R., Gordon, A., & Lavandera, E. (2025, November 6). *"You're not rushing. You're just ready:' Parents say ChatGPT encouraged son to kill himself*. CNN. https://www.cnn.com/2025/11/06/us/openai-chatgpt-suicide-lawsuit-invs-vis

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and "the dark side" of AI. *European Journal of Information Systems: An Official Journal of the Operational Research Society*, *31*(3), 257–268. https://doi.org/10.1080/0960085x.2022.2026621

Minkkinen, M., Zimmer, M. P., & Mäntymäki, M. (2023). Co-shaping an ecosystem for responsible AI: Five types of expectation work in response to a technological frame. *Information Systems Frontiers: A Journal of Research and Innovation*, *25*(1), 103–121. https://doi.org/10.1007/s10796-022-10269-2

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Otto. (2021, September 16). *Ethik vs. Wirtschaftlichkeit, können wir KI vertrauen? | MAIN Session – OTTO [Ethics vs. economics, can we trust AI?]*. Youtube. https://www.youtube.com/watch?v=aiay2hfDiOg

Polanyi, M. (1962). The Republic of science: Its political and economic theory. *Minerva*, *1*(1), 54–73. https://doi.org/10.1007/bf01101453

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. In *arXiv [eess.AS]*. arXiv. http://arxiv.org/abs/2212.04356

Rufo, Y. (2025, July 27). What Guess's AI model in Vogue means for beauty standards. *BBC News*. https://www.bbc.com/news/articles/cgeqe084nn4o

Ryan, M., & Stahl, B. C. (2021). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information Communication and Ethics in Society*, *19*(1), 61–86. https://doi.org/10.1108/jices-12-2019-0138

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Neural Information Processing Systems*, *abs/2205.11487*, 36479–36494. https://doi.org/10.48550/arXiv.2205.11487

Sobek, T., & Montag, J. (2018). Proportionality Test. In *Encyclopedia of Law and Economics* (pp. 1–5). Springer New York. https://doi.org/10.1007/978-1-4614-7883-6_721-1

Stahl, B. C. (2012). Responsible research and innovation in information systems. *European Journal of Information Systems: An Official Journal of the Operational Research Society*, *21*(3), 207–211. https://doi.org/10.1057/ejis.2012.19

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review*, *20*(3), 571. https://doi.org/10.2307/258788

Suddaby, R., Bitektine, A., & Haack, P. (2017). Legitimacy. *Academy of Management Annals*, *11*(1), 451–478. https://doi.org/10.5465/annals.2015.0101

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

Watson, D. S., Mökander, J., & Floridi, L. (2025). Competing narratives in AI ethics: a defense of sociotechnical pragmatism. *AI & Society*, *40*(5), 3163–3185. https://doi.org/10.1007/s00146-024-02128-2

Yilma, K. (2025). From principles to process: the principlist approach to AI ethics and lessons from Internet bills of rights. *AI and Ethics*, *5*(4), 4279–4291. https://doi.org/10.1007/s43681-025-00719-3

Yin, R. K. (1981). The Case Study Crisis: Some Answers. *Administrative Science Quarterly*, *26*(1), 58–65. https://doi.org/10.2307/2392599

Yin, R. K. (1994). *Case Study Research: Design and Methods*. SAGE Publications.

Zhang, Z., & Wang, J. (2024). Can AI replace psychotherapists? Exploring the future of mental health care. *Frontiers in Psychiatry*, *15*, 1444382. https://doi.org/10.3389/fpsyt.2024.1444382

Simon Sturm, M.A. HSG, is a Research Associate and a PhD candidate at the Institute of Information Systems and Digital Business at the University of St.Gallen, Switzerland. His research focuses on the management and governance of machine learning-based AI innovation in electronic commerce.

Address: University of St.Gallen, Institute of Information Systems and Digital Business (IWI-HSG), Müller-Friedberg-Strasse 8, 9000 St.Gallen, Switzerland,
E-Mail: simongabriel.sturm@unisg.ch
ORCID: https://orcid.org/0009-0008-9670-2547

Florian Krause, Dr., is a PostDoc and Lecturer at the Institute for Business Ethics at the University of St.Gallen, Switzerland and at the Institute for Interdisciplinary Work and Employment Studies at the Leibniz University of Hanover, Germany. His research addresses ethics of digital transformation as well as foundational questions in economics and business ethics, with a particular interest in their normative underpinnings.

Address: University of St.Gallen, Institute for Business Ethics (IWE-HSG), Blumenbergplatz 9, 9000 St.Gallen, Switzerland, E-Mail: florian.krause@unisg.ch
ORCID: https://orcid.org/0000-0002-9981-165X

Benjamin van Giffen, Prof. Dr., is an Associate Professor at the University of Liechtenstein, Liechtenstein and an associate researcher with the European Research Center for Information Systems (ERCIS). His research focuses on the organizational adoption of artificial intelligence, AI business value, digital platforms and ecosystems, and human-centered design innovation (e.g., Design Thinking). Prior to joining academia, he held various roles in IT project and compliance management at a global pharmaceutical manufacturing company, where he also served as a manager of the sales and operations planning processes for six sales regions and five manufacturing plants in Europe.

Address: University Liechtenstein, Information Systems and Digital Innovation, Fürst-Franz-Josef-Strasse, 9490 Vaduz, Liechtenstein, E-Mail: benjamin.vangiffen@uni.li
ORCID: https://orcid.org/0000-0001-6753-8399

Received: 28.08.2025 | Revised: 31.10.2025 | Accepted: 09.01.2026

Artificial Intelligence as a Socio-Economic Dilemma: Ordonomic Diagnosis–Reflection–Design for Education, Work and Governance

Patrick Hedfeld

Abstract: This paper analyses artificial intelligence (AI) through the lens of *ordonomics*, a normative-institutional approach that connects economic rationality with ethical reflection. While most discussions in AI ethics focus on principles such as fairness, transparency, and accountability, fewer studies address how these principles can be institutionalized through incentive-compatible rules. We therefore conceptualize AI not as a primarily technological challenge but as a social order problem that requires institutional design and governance. The paper explicitly maps the classical ordonomic three-level schema—actor, institutional order, and market/discourse—onto an applied heuristic of Diagnosis–Reflection–Design, demonstrating how this triad operationalizes ordonomic reasoning for the AI context. Building on this foundation, we identify and categorize key AI-related social dilemmas (economic, epistemic, ethical, and educational). The analysis develops differentiated responsibilities across levels of coordination and proposes rule-based cooperation solutions that align individual incentives with collective welfare. By linking ordonomics to current frameworks such as *Responsible AI*, *algorithmic accountability*, and the *EU AI Act*, the paper positions ordonomics as a design-oriented ethics that bridges normative ideals and institutional economics. The result is a framework for diagnosing conflicts, reflecting responsibilities, and designing cooperative solutions that reconcile innovation with social responsibility.

Keywords: Ordonomics; social dilemmas; artificial intelligence; Diagnosis–Reflection–Design; moral paradox; institutional design; AI governance

Künstliche Intelligenz als sozioökonomisches Dilemma: Eine ordonomische Diagnose–Reflexion–Design Struktur für Bildung, Arbeit und Governance

Zusammenfassung: Dieser Artikel analysiert Künstliche Intelligenz (KI) aus der Perspektive der Ordonomik, einem normativ-institutionellen Ansatz, der ökonomische Rationalität mit ethischer Reflexion verbindet. Während sich die meisten Diskussionen in der KI-Ethik auf Prinzipien wie Fairness, Transparenz und Rechenschaftspflicht konzentrieren, befassen sich weniger Studien mit der Frage, wie diese Prinzipien durch anreizkompatible Regeln institutionalisiert werden können. Wir konzeptualisieren KI daher nicht als primär technologische Herausforderung, sondern als ein Problem der sozialen Ordnung, das institutionelle Gestaltung und Governance erfordert. Der Artikel bildet das klassische ordonomische Drei-Ebenen-Schema – Akteur, institutionelle Ordnung und Markt/Diskurs – explizit auf eine angewandte Heuristik aus Diagnose, Reflexion und Design ab und zeigt, wie diese Triade ordonomisches Denken im KI-Kontext operationalisiert. Aufbauend auf

dieser Grundlage identifizieren und kategorisieren wir zentrale KI-bezogene soziale Dilemmata (wirtschaftliche, epistemische, ethische und pädagogische). Die Analyse entwickelt differenzierte Verantwortlichkeiten auf verschiedenen Koordinationsebenen und schlägt regelbasierte Kooperationslösungen vor, die individuelle Anreize mit dem Gemeinwohl in Einklang bringen. Durch die Verknüpfung der Ordonomik mit aktuellen Rahmenwerken wie Responsible AI, algorithmische Rechenschaftspflicht und dem EU-AI-Act positioniert das Papier die Ordonomik als designorientierte Ethik, die normative Ideale und institutionelle Ökonomie verbindet. Das Ergebnis ist ein Rahmenwerk zur Diagnose von Konflikten, zur Reflexion von Verantwortlichkeiten und zur Entwicklung kooperativer Lösungen, die Innovation mit sozialer Verantwortung in Einklang bringen.

Stichwörter: Ordonomik; soziale Dilemmata; Künstliche Intelligenz; Diagnose–Reflexion–Design; Moralparadox; institutionelles Design; KI-Governance

1. Introduction

The rapid development of artificial intelligence (AI) represents one of the most profound transformation processes of the twenty-first century. Machine learning, generative language models, and algorithmic decision systems not only alter production and service processes but also intervene deeply in education, labor markets, and democratic communication. These developments raise questions that extend far beyond technology: What happens to employment, to equal opportunity, and to the legitimacy of decisions when learning and reasoning are delegated to machines? How can societies ensure that innovation strengthens, rather than erodes, the social foundations on which it depends?

This paper starts from the assumption that AI is not primarily a technical phenomenon, but a problem of institutional order—a problem of how rules, incentives, and norms coordinate the relationship between individual benefit and collective well-being. Approaching AI from this institutional perspective allows us to go beyond moral appeals to responsible behavior and instead analyse the structural incentive constellations that shape action. Many current debates in AI ethics still address dilemmas as if they could be solved through awareness or virtue. Ordonomics offers a complementary approach: it understands moral conflicts as coordination failures that must be resolved through the redesign of social rules.

1.1 Research gap and motivation

Existing research on AI ethics and governance provides a rich set of principles and guidelines—for example, transparency, fairness, accountability, and human oversight (Floridi & Cowls, 2019; Jobin, Ienca & Vayena, 2019). Yet these frameworks often remain declarative rather than institutionalized. They articulate what ought to be valued, but not how competing incentives can be realigned to make responsible behavior rational for the actors involved. Ordonomics uniquely links moral reasoning to incentive design (Pies, 2017a; Pies, 2017b).

This lack of institutional-ethical analysis constitutes a gap in the current debate. The present paper addresses this gap by applying ordonomic reasoning—a framework within normative institutional economics that studies how moral norms and economic incentives can be integrated through rule design. Ordonomics thereby offers a grammar of argumen-

tation that explains social dilemmas as unintended consequences of rational action and seeks cooperative reforms rather than moral condemnation.

1.2 Conceptual orientation: ordonomics as applied framework

While ordonomics is traditionally formulated in terms of three levels of social coordination—

(1) the actor level (rule-following),
(2) the institutional order (rule-setting), and
(3) the market or discourse (rule-finding)—

This triad does not replace the classical scheme but serves as its applied heuristic translation:

Diagnosis corresponds to analysing incentive structures and identifying coordination failures.

Reflection assigns differentiated responsibilities across the actor, institutional, and market-discourse levels.

Design develops rule-based cooperation solutions that transform moral conflicts into win–win constellations through outbidding arguments (Pies, 2017a, Pies 2017b, Pies 2025).

Through this mapping, ordonomics becomes practically applicable for AI governance, offering both theoretical rigor and analytical clarity.

1.3 Positioning within AI ethics and governance

This paper situates ordonomics among established frameworks of Responsible AI, socio-technical systems theory, and algorithmic accountability.

Responsible AI emphasizes ethical principles and value alignment. Socio-technical approaches highlight the interplay of technical artefacts and social structures. Algorithmic accountability introduces procedural mechanisms like audits and documentation.

Ordonomics complements these approaches by focusing on the institutional preconditions under which moral principles become sustainable in practice. It treats fairness, transparency, and accountability not as behavioral exhortations but as rule properties that can be designed, enforced, and competitively rewarded.

1.4 Aim and contribution

The overall aim of this paper is to develop a systematic ordonomic analysis of artificial intelligence as a socio-economic dilemma. Specifically, the contribution is fourfold:

It clarifies the theoretical foundations of ordonomics and maps the classical three-level schema onto the applied triad Diagnosis–Reflection–Design. It provides a structured typology of AI-related social dilemmas—economic, epistemic, ethical, and educational—illustrating how individually rational strategies lead to collective inefficiencies. It elaborates differentiated responsibilities across actor, institutional, and market levels, addressing the moral paradox of modernity, in which individuals are morally overburdened and citizens politically underchallenged. It develops cooperative design strategies—regulatory, organizational, and educational—that align self-interest with the common good, illustrated through case studies on AI in employment, education, and governance.

1.5 Normative orientation and structure

The study is normative and design-oriented. It does not merely describe dilemmas but formulates institutional recommendations that can mitigate them. Accordingly, the paper is structured as follows: Section 2 explains the ordonomic framework and its conceptual mapping. Section 3 diagnoses and categorizes AI-related dilemmas. Section 4 applies the triad in depth, combining theoretical reasoning with practical illustration. Sections 5–7 develop case studies, discussion, and implications for governance, education, and policy.

Through this structure, the paper aims to demonstrate that ordonomics provides a coherent and operational method for linking moral reasoning with institutional design—an approach particularly needed in the governance of artificial intelligence.

2. Theoretical Framework: Ordonomics and the Moral Paradox of Modernity

Ordonomics, as developed primarily by Ingo Pies and colleagues since the 1990s, belongs to the broader tradition of normative institutional economics. It builds on Walter Eucken's classic dictum that "people must not be required to do what the economic order alone can achieve" (Eucken, 1952/1990, p. 368). In this sense, ordonomics conceives of morality not as an alternative to economics, but as a coordination resource that becomes productive through appropriately designed rules, procedures, and institutions (Homann & Pies, 2000). It thereby continues the ordoliberal tradition of German economic thought, combining ethical reasoning with regulatory policy (Ordnungspolitik).

2.1 The ordonomic logic of coordination

Analytically, ordonomics distinguishes three levels of social coordination:

(1) the actor level, where individuals and organizations pursue their interests under given rules;
(2) the institutional level, where rules and procedures are formulated and modified; and
(3) the market or discourse level, where ideas and competition generate incentives for further rule evolution.

This three-level schema provides a systematic architecture for analyzing moral and social conflicts. It reveals how individually rational strategies, when aggregated, can generate collectively inefficient outcomes—a pattern that ordonomics interprets as a social dilemma (Pies, 2000; Beckmann & Pies, 2016; Buchanan, 2000). In such cases, moral exhortations to individuals often fail, because the incentive structures that shape behavior remain unchanged. Instead, ordonomics calls for rule reforms that realign incentives so that self-interest and the common good coincide.

2.2 The moral paradox of modernity

This insight underlies what Pies (2022) describes as the moral paradox of modernity: modern societies tend simultaneously to morally overburden individuals and to politically underchallenge citizens. In other words, systemic coordination failures are often reinterpreted as moral deficits of individuals. For example, the public debate might blame users, consumers, or workers for outcomes that are structurally determined by flawed institu-

tional incentives. The result is a category error: problems of rule design are treated as problems of personal virtue.

Ordonomics resolves this paradox by analytically separating the level of rules (where solutions are designed) from the level of actions (where outcomes occur). This separation allows moral expectations to be rechanneled into institutional arrangements that reward cooperation instead of presupposing it. By doing so, ordonomics transforms moral appeals into rationally enforceable mechanisms of cooperation.

2.3 Social dilemmas as analytical core

Methodologically, ordonomics interprets many social problems as dilemmas of cooperation. In these situations, individually rational strategies generate collectively inferior outcomes—such as environmental degradation, social inequality, or algorithmic bias. The ordonomic response is to generate "outbidding arguments" (Pies, 2017a, Pies 2017b): arguments that make cooperative rule change not only morally desirable but privately attractive to the actors involved. This mechanism ensures that reforms are incentive-compatible rather than coercive or moralistic. The analytic procedure thus proceeds in three interrelated steps—Diagnosis, Reflection, and Design—which correspond directly to the classical three levels of coordination: Diagnosis clarifies the conflict structure and identifies perverse incentives (actor level). Reflection allocates differentiated responsibilities for reform across individual, institutional, and market-discourse levels. Design develops new rules, procedures, and incentives that transform destructive conflicts into mutually beneficial cooperation.

2.4 Ordonomics as normative institutional economics

In its philosophical foundations, ordonomics shares assumptions with the constitutional economics of Buchanan (2000) and the discourse ethics of Habermas (1981): both seek to reconcile rational self-interest with moral legitimacy through rule-based cooperation. Yet ordonomics distinguishes itself by treating moral norms explicitly as resources for coordination, not merely as constraints. Its focus lies on institutional learning processes—how societies improve their rules by identifying and correcting structural inefficiencies (Minnameier, 2016).

In this way, ordonomics transcends the traditional opposition between morality and economics. Rather than judging behavior from a moral standpoint, it examines how rules can make moral behavior rational. This shift from appeal to design marks ordonomics as a practical, action-guiding framework for ethics in modern, complex societies.

2.5 Application to artificial intelligence

Applied to artificial intelligence, the ordonomic perspective produces two key insights.

First, it highlights where the real coordination failures occur: in the rules, incentives, and information structures that govern data use, algorithmic decision-making, and accountability (Selbst et al., 2019; Raji et al., 2020). Second, it cautions against the moralization of AI discourse—for instance, when public debates focus on "ethical AI" as a matter of individual responsibility, while neglecting the institutional and regulatory dimensions.

Ordonomics provides a lens to differentiate responsibilities:

- What competences are required at the individual level?
- What rules and procedures must institutions establish?
- What market and discourse structures ensure that private interest aligns with the public good?

Through these guiding questions, the framework prevents the reduction of complex socio-technical systems to individual virtue ethics and instead treats AI as a governance problem. In this sense, ordonomics bridges normative theory, institutional design, and economic reasoning—making it particularly well suited to the ethical challenges of artificial intelligence.

Ordonomic level	Analytical focus	Guiding questions (examples)	Applied Triad
Actor level	Individual and organizational behaviour	What actions are rational under existing rules?	Diagnosis
Institutional level	Rules, procedures, governance structures	How are incentives shaped and responsibilities allocated?	Reflection
Market / discourse level	Public debate, norm evolution	How can rules be reformed through cooperation and new technology?	Design

Table 1. Mapping the classical ordonomic levels of coordination onto the applied heuristic of Diagnosis–Reflection–Design.

The schematic in this table visualizes how the classical ordonomic levels of coordination (actor, institutional, and market/discourse) correspond to the applied triad of Diagnosis–Reflection–Design. It illustrates the process logic by which AI-related social dilemmas—economic, epistemic, ethical, and educational—are diagnosed, reflected, and institutionally redesigned toward cooperative outcomes.

A conceptual framework mapping the classical ordonomic three-level schema (actor, institutional, and market–discourse levels) onto the applied heuristic of Diagnosis–Reflection–Design in the analysis of AI-related social dilemmas.

3. Artificial Intelligence as a Transformation Phenomenon and Source of Socio-Economic Dilemmas

Artificial intelligence represents not only a technological innovation but a profound transformation of social coordination. As with earlier general-purpose technologies such as electricity or the Internet, AI reconfigures the way information, labor, and responsibility are distributed within society. Yet unlike previous industrial revolutions, AI operates at a cognitive level: it automates not only physical tasks but also judgment, evaluation, and decision-making. This dual nature—technological and epistemic—makes AI a paradigmatic case for ordonomic analysis. It does not create entirely new moral questions but intensifies existing social dilemmas by scaling them through automation, data, and network effects.

3.1 Defining the transformation phenomenon

AI can be described as a transformation phenomenon in two interconnected senses.

First, it transforms productive structures by automating and optimizing tasks that were previously human. Second, it transforms institutional structures, altering how knowledge, authority, and legitimacy are distributed. The technology therefore acts simultaneously on the economic and the normative infrastructures of society. This dual impact leads to a constellation of social dilemmas—conflicts between rational individual optimization and collective welfare—that require not moral exhortation but institutional realignment.

Ordonomics conceptualizes such transformation processes as rule dynamics: new technologies challenge the adequacy of existing rules, which then triggers discursive and institutional adaptation. The challenge for AI governance is to ensure that this adaptation process does not lag behind technological progress but evolves in parallel, maintaining the compatibility between innovation and cooperation.

3.2 Typology of AI-related social dilemmas

To analyze these coordination challenges, AI-related dilemmas can be grouped into four interrelated categories: economic, epistemic, ethical, and educational. Each category captures a distinctive misalignment of incentives that arises when private rationality diverges from collective welfare.

(a) Economic dilemmas: efficiency and employment

At the firm level, AI promises efficiency gains, cost reduction, and competitive advantage. Rational companies therefore invest in automation, data analytics, and algorithmic management. Yet when aggregated across the economy, these micro-level decisions may generate macro-level side effects: the polarization of skill profiles, displacement of middle-income jobs, and rising inequality (Acemoglu & Restrepo, 2020). This tension between individual competitiveness and collective stability constitutes a classical social dilemma.

From an ordonomic viewpoint, the question is not whether automation is good or bad, but under which institutional conditions its benefits can be distributed without eroding social cohesion. Rules for retraining, social insurance, and innovation incentives determine whether the dilemma remains functional (promoting progress) or becomes dysfunctional (undermining solidarity). Written in other terms:

- Actors: Firms competing in product and labor markets invest in AI-driven automation to improve productivity and reduce costs.
- Incentives: Competitive pressure rewards early adoption, efficiency gains, and short-term cost reduction, making automation a rational firm-level strategy.
- Collective outcome: When aggregated, these individually rational decisions contribute to job polarization, the displacement of middle-income work, and rising inequality.
- Ordonomic insight: The resulting tension between firm-level efficiency and macro-level social stability constitutes a social dilemma that can only be addressed through institutional rules for reskilling, social insurance, and labor market adaptation.

(b) Epistemic dilemmas: performance and transparency

AI systems often trade accuracy for explainability. Proprietary algorithms, trained on vast datasets, outperform humans in prediction but resist scrutiny. The resulting opacity generates what Burrell (2016) calls the black-box problem. Companies have an incentive to protect intellectual property; regulators and citizens demand transparency and accountability. The conflict lies between the private incentive to conceal and the public need to understand.

Ordonomically, this dilemma is not solved by appealing to corporate ethics but by designing institutions—such as audit rights, model documentation (Gebru et al., 2018), and disclosure standards—that make transparency compatible with competitive interest.

- Actors: AI developers and deploying organizations on the one hand, regulators and affected publics on the other.
- Incentives: Firms seek to protect proprietary models and competitive advantage, while regulators and citizens require transparency to ensure accountability.
- Collective outcome: The resulting opacity undermines trust and democratic oversight, despite high system performance.
- Ordonomic insight: Transparency must be institutionalized through audit rights and documentation standards that make accountability incentive-compatible.

(c) Ethical dilemmas: optimization and fairness

Algorithmic optimization frequently reproduces structural bias present in data (Binns, 2018; Raji et al., 2020). Employers, insurers, or credit institutions pursue efficiency and risk minimization; in aggregate, such optimization can undermine fairness and social legitimacy. The dilemma arises when private utility maximization erodes the collective basis of trust.

The ordonomic lens interprets fairness as a property of rules, not of isolated actions. The ethical challenge is therefore to reform the rule systems that govern data collection, model training, and evaluation—turning fairness from a moral aspiration into an institutional standard.

- Actors: Organizations deploying AI systems for hiring, credit scoring, insurance, or public administration, as well as individuals affected by algorithmic decisions.
- Incentives: Organizations are incentivized to optimize decision accuracy, efficiency, and risk minimization, often relying on historical data and automated optimization criteria.
- Collective outcome: When scaled across systems and institutions, such optimization can reproduce or amplify existing social biases, undermining fairness, equal opportunity, and the legitimacy of algorithmic decision-making.
- Ordonomic insight: Fairness deficits are not primarily the result of unethical individual behavior but of rule systems that reward efficiency without internalizing distributive effects; addressing this dilemma therefore requires institutional standards for data governance, bias auditing, and accountability that make fairness incentive-compatible.

(d) Educational dilemmas: personalization and autonomy

In education, AI enables personalized learning and efficient assessment. Adaptive systems can tailor tasks, feedback, and pacing to each student (Seufert & Meier, 2023). Yet such

personalization risks eroding students' autonomy if it replaces self-regulation with algorithmic steering. The dilemma here is between short-term learning gains and long-term independence.

When teachers rely excessively on AI feedback or grading, the epistemic authority shifts from human educators to technical systems. The ordonomic approach suggests designing rules that preserve the normative primacy of human judgment, for instance through transparency obligations, mixed-assessment formats, and digital literacy education.

- Actors: Students, teachers, educational institutions, and providers of AI-based learning and assessment systems.
- Incentives: AI systems offer incentives for efficiency, personalization, and performance optimization by adapting content, feedback, and assessment to individual learners.
- Collective outcome: When reliance on algorithmic guidance becomes pervasive, educational practices risk prioritizing measurable performance over critical reflection, thereby weakening learners' autonomy, epistemic agency, and responsibility for their own learning processes.
- Ordonomic insight: Preserving educational autonomy requires rules that embed AI as a supportive instrument rather than a substitutive authority; institutional designs such as transparency requirements, mixed assessment formats, and AI literacy education can realign efficiency gains with the public good of independent judgment.

3.3 Cross-sectional synthesis

These four domains reveal recurring structural patterns. Across them, AI generates coordination failures by altering information asymmetries, incentive structures, and moral expectations simultaneously. Actors optimize locally within their constraints, while collectively undermining the institutional trust that sustains cooperation.

From an ordonomic perspective, these tensions can be summarized as three generic conflicts:

- Information asymmetry – AI systems concentrate information and decision power asymmetrically between producers and users.
- Incentive misalignment – rules reward short-term efficiency rather than long-term stability or fairness.
- Moral overburdening – responsibility is displaced onto individuals who lack the power to alter the underlying incentives.

The consequence is a proliferation of what Pies (2017) calls undesirable dilemmas—those that trap actors in collectively inferior equilibria. Recognizing these structures is the first step toward diagnostic clarity. The next analytical move is reflection: to assign differentiated responsibilities across actors, institutions, and the market-discourse system, preparing the ground for cooperative redesign.

3.4 Interim conclusion

AI thus functions as a mirror for the social order. It exposes latent contradictions between efficiency, justice, and legitimacy that modern societies must address not by moralizing technology but by improving their rules of coordination. The ordonomic approach provides a systematic method to analyse these contradictions: it treats AI neither as an

autonomous moral agent nor as a neutral tool, but as an amplifier of human and institutional incentives. The following section applies this ordonomic triad—Diagnosis, Reflection, and Design—to develop concrete pathways for institutional adaptation in the domains of work, education, and governance.

4. Applying the Ordonomic Triad (Diagnosis–Reflection–Design)

Building on the theoretical foundations and the typology of dilemmas, this section applies the ordonomic triad to the domain of artificial intelligence. The triad serves as a heuristic that connects empirical observation with institutional design. It enables a movement from diagnosis—the clarification of conflict structures—to reflection—the assignment of differentiated responsibilities—and finally to design, the development of cooperative solutions that transform dysfunctional equilibria into socially beneficial ones.

The application demonstrates that AI governance, when viewed through an ordonomic lens, is not primarily about restricting innovation, but about shaping the rules that align innovation with collective welfare.

4.1 Diagnosis: AI and employment as an illustrative dilemma

The employment dilemma illustrates the ordonomic logic of social conflicts in a paradigmatic way. On the actor level, companies rationally pursue automation and algorithmic management to increase productivity, flexibility, and competitiveness. The resulting technological substitution effects—especially in repetitive and information-processing tasks—improve firm efficiency and shareholder returns. However, at the systemic level, the same optimization process can produce side effects such as job polarization, regional inequality, and declining middle-class stability (Acemoglu & Restrepo, 2020).

This tension between micro-rationality and macro-irrationality constitutes a classic social dilemma. Each firm acts rationally under competition, but collectively these actions may erode the very demand and trust conditions that sustain markets. In the short term, automation seems welfare-enhancing; in the long term, it risks social fragmentation. Ordonomically, this reveals a coordination failure: rules at the institutional level insufficiently internalize the social costs of technological displacement.

Whether this dilemma is desirable or undesirable depends on the institutional framework. Desirable dilemmas drive progress by rewarding innovation while maintaining adaptive institutions. Undesirable dilemmas, by contrast, persist when institutions fail to redistribute the gains or provide adjustment mechanisms. The ordonomic diagnosis, therefore, distinguishes between functional competition and dysfunctional erosion, identifying the rule conditions that determine the balance.

To exemplify this, consider the differing trajectories of AI diffusion across economies. Countries with strong vocational education systems, portable social insurance, and proactive retraining policies—such as Denmark or the Netherlands—are better equipped to absorb technological shocks without amplifying inequality. In contrast, where institutions rely primarily on labor market deregulation and limited social buffering, the same innovation dynamic produces exclusion rather than empowerment. The dilemma thus manifests not in the technology itself, but in the rules that govern adaptation.

4.2 Reflection: Assigning responsibilities across levels of coordination

Following the ordonomic logic, effective resolution of social dilemmas requires a clear differentiation of responsibilities across actors, institutions, and the market-discourse system.

At the actor level, firms and individuals are expected to act within given rules, but not to unilaterally solve systemic conflicts. The expectation that corporations alone should "act ethically" disregards the competitive pressure that constrains them. However, actors can contribute by integrating AI ethics principles into internal governance, promoting transparency in algorithmic systems, and supporting retraining initiatives for affected employees. These actions mitigate, but do not eliminate, the structural causes of the dilemma.

At the institutional level, governments, regulators, and intermediary organizations must establish the formal and informal rules that correct incentive asymmetries. This includes labor market institutions that facilitate reskilling and job mobility, fiscal systems that support social innovation, and regulatory frameworks that ensure accountability for AI deployment in employment contexts. Educational institutions have a parallel role: integrating AI literacy into curricula and ensuring that learning systems preserve critical reasoning rather than promote dependency.

At the market and discourse level, the competitive and communicative arenas in which norms evolve, public debate plays a constitutive role. The recent open letters on AI governance—whether calling for a moratorium on large-scale experiments (Future of Life Institute, 2023) or for a pause in overregulation (AI Champions, 2025)—illustrate how the discourse level functions as a meta-arena of rule-finding. In this space, societies negotiate the legitimate boundaries between innovation and precaution. Discursive coordination, however, requires procedural rationality: actors must learn to "outbid" purely moralistic arguments by demonstrating that cooperation is instrumentally superior to unilateral action.

This reflective differentiation avoids two extremes: it neither moralizes market behavior nor reduces ethics to compliance. Instead, it conceptualizes social responsibility as a distributed task, embedded in complementary roles that together enable cooperative rule change. The moral paradox of modernity is thus countered: individuals are no longer overburdened, and citizens are not underchallenged, because each level is normatively engaged within its functional competence.

4.3 Design: From moral appeal to rule-based cooperation

The final step of the triad, Design, focuses on developing institutions and mechanisms that realign incentives. The ordonomic ambition is not to impose external constraints on rational actors but to make cooperation rational by transforming the rule environment. In the AI context, this entails the creation of regulatory, organizational, and educational designs that embed normative goals within competitive structures.

Regulatory design addresses the systemic level of rule-setting. The emerging EU Artificial Intelligence Act exemplifies a move toward risk-based governance, where AI systems are classified according to their societal impact. Requirements for transparency, auditability, and human oversight translate moral concerns into operational criteria. Rather than restricting innovation, such regulation creates a predictable playing field in which ethical performance becomes a competitive advantage. By coupling compliance incentives with public accountability, these rules internalize externalities without centralizing control.

Organizational design operates at the meso-level of companies and institutions. Mechanisms such as algorithmic impact assessments, model cards (Mitchell et al., 2019), and data documentation protocols (Gebru et al., 2018) institutionalize accountability. They make the ethical quality of AI systems observable and auditable, turning trust into a measurable asset. Firms that adopt such measures not only reduce reputational risk but also signal reliability to clients and regulators. Over time, these practices generate a market for responsibility, where transparency and fairness are rewarded.

Educational design complements these mechanisms by addressing the formation of competences and moral agency. Integrating AI literacy into education helps future workers, managers, and citizens understand the logics and limits of algorithmic systems. Rules for AI use in schools and universities should protect intellectual autonomy, for example by combining automated feedback with oral defense and reflective documentation. In doing so, education becomes a microcosm of ordonomic reasoning: it aligns individual learning with collective epistemic responsibility.

Across these design levels, the principle of outbidding arguments serves as the connective logic. Actors are motivated to cooperate when the institutional framework allows them to achieve more together than apart. Cooperation becomes rationally superior to defection because well-designed rules transform moral desirability into private advantage. In this sense, ordonomic design is not an abstract ideal but a strategy of institutional evolution—a process of iterative rule improvement through public reasoning and competitive experimentation.

4.4 Synthesis: Toward a cooperative order of AI governance

Applying the ordonomic triad to AI governance reveals that the key to reconciling innovation and responsibility lies in rule-based cooperation. The dilemmas of employment, transparency, and autonomy cannot be solved by technical optimization alone. They require institutional arrangements that transform the structure of incentives and the distribution of knowledge.

The ordonomic contribution is to offer a methodological grammar for such institutional design. Diagnosis identifies the coordination failures; reflection clarifies the normative division of labor; and design formulates concrete mechanisms that realign private and public interests. When applied iteratively, this process enables what Homann (2002) described as the "ethical learning of systems"—a collective capacity to adapt rules in light of new technological realities.

In the domain of artificial intelligence, this means that responsibility is not exhausted in compliance checklists or moral declarations. It resides in the continuous improvement of the institutional order itself—the dynamic adjustment of rules that make innovation sustainable. Through this perspective, ordonomics reframes AI ethics as a project of institutional creativity, in which societies learn to design the conditions under which human and artificial intelligence can coexist productively.

5. Case Studies: Education, Human Resources, and Governance

The ordonomic framework becomes tangible when applied to concrete domains in which artificial intelligence reshapes established coordination patterns. Each domain presents specific constellations of actors, incentives, and rule systems, yet they all reflect the same

underlying logic: individually rational optimization that unintentionally produces collective dysfunction. Analysing these cases through the triad Diagnosis–Reflection–Design demonstrates how cooperation can be reconstructed by institutional means.

5.1 Education: Learning under algorithmic conditions

Diagnosis: In education, AI tools such as adaptive tutoring systems, plagiarism detectors, and generative language models have transformed both teaching and assessment. Students use AI to obtain feedback and accelerate learning, while teachers experiment with automation to manage workloads and personalize instruction. On the actor level, these behaviors appear rational: they promise efficiency, accessibility, and inclusion. Yet at the systemic level, they risk undermining the cultivation of critical thinking, authorship, and intellectual autonomy.

Reflection: The resulting dilemma lies in the tension between personalization and autonomy. Algorithmic systems optimize for engagement and performance metrics; human learning, however, depends on friction, error, and reflection. When algorithms overfit education to predicted outcomes, they silently redefine what counts as competence. This shift threatens the public good of education as a space for independent reasoning (Williamson & Piattoeva, 2022). From an ordonomic standpoint, the task is to design rules that preserve the normative primacy of human judgment within digitally mediated learning environments. Reflection therefore differentiates responsibilities: teachers remain accountable for assessment standards; institutions set transparent usage policies for AI tools; students are obliged to disclose AI assistance; and ed-tech providers ensure traceability and explainability of their systems.

Design: One can translate these normative expectations into practice. Universities may require written process documentation or oral defense to accompany AI-supported assignments; curricula may include modules on digital epistemology and algorithmic bias; accreditation agencies may demand audit trails for automated grading. Through such institutional embedding, AI becomes not an agent of substitution but an instrument of cognitive cooperation—enhancing learning while maintaining the rule of reflective autonomy (Minnameier, 2025).

5.2 Human resources: Algorithmic management and the search for fairness

Diagnosis: The domain of human resource management illustrates another structural dilemma. Organizations increasingly rely on AI-based tools for recruitment, screening, and performance evaluation. These systems promise objectivity, speed, and cost efficiency, allowing firms to handle thousands of applications or to monitor productivity in real time. For each company, adopting such technologies appears strategically rational; no single actor can afford to ignore efficiency pressures. However, the aggregate effect of widespread automation may be the erosion of fairness and trust in labor markets (Raghavan et al., 2020).

The incentive asymmetry is evident because vendors profit from proprietary algorithms, employers from efficiency, while job seekers bear the opacity and potential bias of automated judgments. The collective outcome is a deficit of legitimacy—an institutional externality rather than a moral lapse.

Reflection: The ordonomic analysis reframes this as a rule-design problem, in other words: how can fairness become an economically rational attribute of AI systems? The answer lies in making fairness auditable and commercially valuable. Certification schemes, bias-testing requirements, and transparency clauses in procurement contracts turn moral expectations into enforceable standards. Reflection again distributes tasks means developers document models and datasets; employers provide audit access and impact assessments; regulators define thresholds and procedures for algorithmic accountability.

Design: One could connect these layers through shared metrics—algorithmic impact assessments, model cards, and datasheets for datasets—that create comparability across systems (Gebru et al., 2021; Mitchell et al., 2019). When such reporting becomes a market norm, reputational and legal incentives converge, producing what might be called a market for responsibility. In this order, ethical quality is no longer an externality but a competitive resource.

5.3 Governance: Balancing innovation and precaution

Beyond specific sectors, AI challenges the foundations of democratic governance itself.

Diagnosis: Public authorities experiment with predictive policing, welfare automation, and administrative decision systems. These applications raise legitimacy questions that cannot be answered solely by technical accuracy. The governance dilemma concerns the balance between innovation and precaution. Excessive regulation risks stifling beneficial innovation; too little regulation exposes citizens to opaque and potentially discriminatory systems. Diagnosing this dilemma reveals an institutional misalignment: the pace of technological change outstrips the adaptability of legal frameworks. Governments, motivated by efficiency and cost reduction, adopt algorithmic tools faster than oversight mechanisms evolve. Citizens, meanwhile, lack the information and rights necessary to contest automated decisions. The result is a gap between de facto technological power and de jure democratic control.

Reflection: It requires clarifying who bears responsibility for maintaining this balance. Legislators set general principles; agencies implement and monitor; civil society and academia supply critical feedback. Discourse at the public level functions as a corrective meta-game in which legitimacy is continually renegotiated. The rule of law thus depends on procedural transparency and access to contestation.

Design: One could translate these insights into institutional architecture. Risk-based regulation, such as the EU Artificial Intelligence Act, classifies AI systems by societal impact and imposes corresponding duties of transparency, data governance, and human oversight. Complementary instruments—impact assessments, mandatory documentation, and independent audit bodies—transform moral demands for accountability into enforceable rights. Through these mechanisms, governance evolves from reactive control to anticipatory coordination: a system capable of learning from its own dilemmas.

5.4 Comparative synthesis

Across the domains of education, human resources, and governance, the same structural pattern reappears. AI systems intensify coordination failures by shifting information asymmetries and by multiplying the speed of interaction, while existing institutions lag behind. Yet these failures are not inevitable. When rules are designed to realign incentives—so that

transparency, fairness, and learning become rationally rewarded—the dilemmas turn from destructive to productive.

Ordonomically, this transformation can be understood as a process of institutional learning. Each domain develops specific mechanisms—usage policies in education, audit standards in HR, risk classification in governance—but all follow the same meta-logic: making cooperation more rewarding than unilateral optimization. In this sense, AI becomes a test case for modern societies' capacity to update their moral infrastructures. Rather than appealing to ethical heroism, ordonomics invites the redesign of the game itself, ensuring that innovation and integrity reinforce one another.

6. Discussion: Integrating Ordonomics into the AI Ethics and Governance Discourse

The preceding analysis demonstrates that artificial intelligence confronts societies with classical coordination problems in a new technological guise. The ordonomic framework offers a language and logic for understanding these problems not as isolated ethical controversies but as manifestations of incentive misalignment. This section situates ordonomics within the broader landscape of AI ethics and governance and discusses its theoretical, methodological, and normative implications.

6.1 Bridging AI ethics and institutional economics

Much of the current AI ethics discourse is principle-driven. Frameworks such as the OECD AI Principles (2019) or Floridi and Cowls' (2019) five-principle model—beneficence, non-maleficence, autonomy, justice, and explicability—formulate shared values but often remain detached from the mechanisms that could realize them. Ordonomics contributes an institutional bridge by asking under which rule conditions these values can become self-enforcing.

Whereas traditional ethics seeks moral compliance by individuals, ordonomics focuses on the game architecture that structures behavior. It conceptualizes cooperation as a product of rule-design rather than virtue. This perspective resonates with the notion of responsible innovation (Stilgoe et al., 2013) and value-sensitive design (Friedman & Hendry, 2019) but adds an explicit economic dimension: actors must have incentives to internalize ethical expectations. By converting moral aims into institutional payoffs—through regulation, market signaling, and discourse—ordonomics operationalizes what otherwise remains aspirational.

6.2 Complementarity with socio-technical and accountability approaches

Socio-technical systems theory interprets technology as embedded in human, organizational, and cultural contexts (Bijker, 1997; Suchman, 2007). Algorithmic accountability research similarly seeks procedural safeguards such as audits, documentation, and contestability (Raji et al., 2020; Selbst et al., 2019). Ordonomics complements these approaches by providing a normative-economic grammar for why such procedures matter.

Audits and transparency reports, for instance, can be read as institutionalized outbidding arguments: they make trust commercially valuable. Documentation and contestability mechanisms serve to re-balance asymmetric information. From an ordonomic viewpoint, these measures are not external ethical add-ons but rule-based corrections that restore cooperation in competitive settings. The approach thus links descriptive analyses of socio-

technical complexity with a theory of institutional learning—how societies adjust rules to turn dilemmas into opportunities for collective gain.

Recent work on implicit decision-making and governance further supports this perspective. Hedfeld (2025) shows how implicit voting mechanisms and language models can be conceptualized as normatively legitimate and institutionally implementable rules, thereby illustrating how discursive coordination can be translated into incentive-compatible governance structures in complex socio-technical systems (Hedfeld, 2025a; Hedfeld 2025b).

6.3 The ordonomic contribution to governance theory

In governance theory, ordonomics clarifies how discursive and regulatory arenas interact. The discourse level—public debate, scientific reflection, and stakeholder dialogue—acts as a meta-arena for rule-finding. Here, actors test and refine competing arguments, selecting those that enable broader cooperation. The market and institutional levels then implement these insights through laws, standards, and organizational routines.

This iterative feedback between discourse and order constitutes a dynamic model of ethical governance. It aligns with reflexive governance theories (Voß et al., 2006) but grounds them in incentive logic. Rather than ideal deliberation, ordonomics assumes bounded rationality and competitive pluralism. Ethical progress occurs when rule systems evolve to make cooperative solutions Pareto-superior to conflictual ones. In AI governance, this principle underlies risk-based regulation, co-regulation, and standardization efforts that allow innovation to continue under clear accountability constraints.

6.4 Relating ordonomics to current policy frameworks

The emerging EU Artificial Intelligence Act operationalizes many ordonomic ideas in practice. By linking risk categories to mandatory obligations—such as transparency, human oversight, and post-market monitoring—it creates a graduated incentive structure that rewards responsible innovation. Similar logics guide the OECD AI Principles and the UNESCO Recommendation on the Ethics of AI. These instruments function as meta-rules: they do not prescribe outcomes but specify procedures for ethical alignment.

Ordonomics interprets these developments as part of a broader rule evolution: societies learn to transform moral expectations into enforceable coordination mechanisms. Regulation thus becomes an instrument of ethical learning rather than a barrier to progress. The ultimate objective is not to moralize technology but to embed its operation in a framework of reciprocal advantages—a cooperative order in which compliance with ethical standards coincides with strategic rationality.

6.5 Theoretical synthesis and implications

Integrating ordonomics into the AI governance discourse yields three conceptual insights:

Reconceptualizing responsibility. Responsibility shifts from individual virtue to the design of institutions that make virtuous action rational. This redefinition dissolves the moral paradox of modernity: actors are no longer expected to sacrifice their interests but to pursue them within fair rules.

Ethics as institutional learning. Ethical norms evolve through feedback between action, order, and discourse. AI governance exemplifies this evolutionary process: public controversies trigger regulatory adaptation, which in turn reshapes incentives and behavior.

Cooperation as a design problem. The central question of AI governance is not what values to hold, but how to design rules so that adherence to those values becomes payoff-consistent. Ordonomics provides a methodological grammar for this transformation, translating moral reasoning into institutional architecture.

Through these implications, ordonomics positions itself as both a complement and a corrective to mainstream AI ethics. It does not replace value frameworks but grounds them economically, ensuring that normative aspirations survive contact with strategic reality. In this sense, ordonomics represents a form of design-oriented ethics: an ethics of systems, not of saints.

7. Conclusion and Outlook

Artificial intelligence is both a catalyst and a mirror of contemporary social order. Its diffusion amplifies long-standing coordination problems between efficiency, fairness, and legitimacy. By applying ordonomic reasoning, this paper has shown that these tensions cannot be resolved through moral appeals alone but require institutional learning.

Ordonomics, understood as normative institutional economics, interprets social dilemmas as structural misalignments between individual incentives and collective outcomes. Through the triad Diagnosis–Reflection–Design, the approach provides a systematic grammar for transforming moral conflicts into cooperation problems that can be solved by redesigning rules.

Diagnostically, AI reveals where incentive structures generate collectively inferior outcomes: in the displacement of labor, the opacity of algorithmic decision-making, or the erosion of educational autonomy. Reflection then assigns differentiated responsibilities across actors, institutions, and the market-discourse system, preventing both moral overburdening and political passivity. Finally, design translates normative expectations into incentive-compatible rules—through regulation, organizational standards, and education—so that responsibility becomes a structural feature of the system itself.

The broader implication is that AI governance represents an ongoing experiment in institutionalized ethics. Risk-based regulation, algorithmic audits, and transparency requirements exemplify how societies transform values into enforceable coordination mechanisms. Rather than opposing innovation and morality, ordonomics invites their integration: it demonstrates how cooperative rule design can make responsibility a competitive advantage. This article itself can be read as a discursive contribution at the rule-finding (meta-meta-game) level as theory-driven conceptual paper.

Future research should empirically investigate how such ordonomic mechanisms operate in practice—how firms, regulators, and educational institutions internalize ethical expectations through incentive structures. Comparative studies across sectors and jurisdictions could reveal patterns of institutional learning that help societies anticipate, rather than merely react to, technological change.

Ultimately, the ordonomic approach reframes AI ethics as a project of collective rule intelligence. It asks not only how machines can become intelligent, but how humans can design the social rules that ensure intelligence—human or artificial—serves the cooperative advancement of society.

References

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. *Journal of political economy*, *128*(6), 2188-2244.

AI Champions. (2025). Stop the Clock: Open Letter Calling for an EU AI Act Pause. Available online at: https://aichampions.eu (Call 16.01.2026)

Beckmann, M., & Pies, I. (2016). The constitution of responsibility: Toward an ordonomic framework for interpreting (corporate social) responsibility in different social settings. In *Order ethics: An ethical framework for the social market economy* (pp. 221-250). Cham: Springer International Publishing.

Bijker, W. E. (1997). *Of bicycles, bakelites, and bulbs: Toward a theory of sociotechnical change*. MIT press.

Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT), 149–159.

Buchanan, J. M. (2000). *Reason of Rules—Constitutional Political Economy*. Liberty Fund Incorporated, us.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, *3*(1), 2053951715622512.

Eucken, W. (1952/1990). Grundsätze der Wirtschaftspolitik. Tübingen: J.C.B. Mohr (Paul Siebeck).

Floridi, L. & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Philosophy & Technology, 32(4), 685–703.

Friedman, B. & Hendry, D. (2019). Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press.

Future of Life Institute. (2023). Pause Giant AI Experiments: An Open Letter. Available online at: (https://futureoflife.org/open-letter/pause-giant-ai-experiments/) (Call 16.01.2026)

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Hedfeld, P. (2025a). Implicit decision voting made by humans as normative and implementable rules with the help of language models. In R. Buchkremer, O. Koch & A. Lischka (Hrsg.), ifid Schriftenreihe: Beiträge zu IT-Management & Digitalisierung (Bd. 3). FOM-Hochschule für Oekonomie & Management. ISBN 978-3-89275-395-7.

Hedfeld, P. (2025b). Essay: Mit der Langfristigkeit im Herzen–Nachhaltigkeit und Generationengerechtigkeit, eine interdisziplinäre Perspektive zwischen Sozialpädagogik und Wirtschaftsethik. *Zeitschrift für Sozialpädagogik*, (1).

Homann, K. (2002). *Vorteile und Anreize: Zur Grundlegung einer Ethik der Zukunft*. Mohr Siebeck.

Homann, K., & Pies, I. (2000). Wirtschaftsethik und Ordnungspolitik–Die Rolle wissenschaftlicher Aufklärung. *Ordnungstheorie und Ordnungspolitik–Konzeptionen und Entwicklungsperspektiven, Stuttgart*, 329-346.

Jobin, A., Ienca, M. & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence, 1(9), 389–399.

Minnameier, G. (2016). Rationalität und Moralität: Zum systematischen Ort der Moral im Kontext von Präferenzen und Restriktionen. *Zeitschrift für Wirtschafts-und Unternehmensethik*, *17*(2), 259.

Minnameier, G. (2025). Ordonomik und Bildung: Verantwortung für die moderne Gesellschaft (p. 372). wbv Media.

Mitchell, M. (2019). Artificial Intelligence: A Guide for Thinking Humans. Farrar, Straus and Giroux.

Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model Cards for Model Reporting. Proceedings of FAT 2019, 220–229.

OECD. (2019). OECD Principles on Artificial Intelligence. Paris: OECD Publishing. https://www.oecd.org/en/topics/ai-principles.html (Call 16.01.2026)

Pies, I. (2000). Ordnungspolitik in der Demokratie: Ein ökonomischer Ansatz diskursiver Politikberatung. Tübingen: Mohr Siebeck. ISBN 3-16-147507-0.

Pies, I. (2017a). Ordonomik als Methode zur Generierung von Überbietungsargumenten: Eine Illustration anhand der Flüchtlings (politik) debatte (No. 2017-03). Diskussionspapier. https://doi.org/10.5771/1439-880X-2017-2-171

Pies, I. (2017b). The ordonomic approach to business ethics. *Available at SSRN 2973614*.

Pies, I. (2022). Kapitalismus und das Moralparadoxon der Moderne. Berlin: wvb Wissenschaftlicher Verlag Berlin. ISBN 978-3-96138-310-8

Pies, I. (2025). The interplay of incentives and ideas: An intellectual journey from order economics through order ethics to ordonomics (No. 2025-08). Diskussionspapier. https://www.econstor.eu/bitstream/10419/325828/1/1936155664.pdf (Call 16.01.2025)

Raghavan, M., Barocas, S., Kleinberg, J. & Levy, K. (2020). Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT), 469–481.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33-44).

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59-68).

Seufert, S., & Meier, C. (2023). Hybrid Intelligence: Collaboration with AI Systems for Knowledge Work. *HMD Praxis der Wirtschaftsinformatik*, *60*(6), 1194-1209.

Stilgoe, J., Owen, R. & Macnaghten, P. (2013). Developing a Framework for Responsible Innovation. Research Policy, 42(9), 1568–1580.

Suchman, L. (2007). Human-Machine Reconfigurations: Plans and Situated Actions. Cambridge University Press.

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO. https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence (Call 16.01.2026)

Voss, J. P., Bauknecht, D., & Kemp, R. (Eds.). (2006). *Reflexive governance for sustainable development*. Edward Elgar Publishing.

Williamson, B., & Piattoeva, N. (2022). Education governance and datafication. *Education and Information Technologies*, *27*, 3515-3531.

Patrick Hedfeld, Dr., works for the Center of Business Ethics at the Goethe University and as lecturer for the FOM University of Applied Sciences in Frankfurt am Main.

Address: Johann Wolfgang Goethe-Universität Frankfurt Theodor-W.-Adorno-Platz 1 60323 Frankfurt am Main and FOM University of Applied Sciences Franklinstraße 52, 60486 Frankfurt am Main, Germany, E-Mail: hedfeld@econ.uni-frankfurt.dehttps://orcid.org/0000-0002-0385-2829

Received: 30.09.2025 | Revised: 08.12.2025 | Accepted: 19.01.2026

Towards AI Governance in DAX40: A Typology of Organizational Guidelines for Self-Regulation

Niklas Obermann, Daniel Lupp, Uta Wilkens

Abstract: In this article, we combine the discourse on the ethical challenges of using Artificial Intelligence (AI) with research perspectives on overarching AI governance. Although AI govenance is not yet institutionalized, an increasing number of organizations are formulating their own guidelines for the responsible use of AI. These self-regulatory approaches serve as guidance for customers and employees. At the same time, they serve to align organizational processes and control mechanisms. Little is known about the differences between self-regulatory approaches and how organizations anticipate the future direction of AI governance. Using DAX40 companies as an example, we examine how organizations design AI guidelines for self-regulation and which criteria of AI ethics they take into account. Based on a systematic search strategy and qualitative content analysis, we identify three different types of self-regulation: (1) non-codified self-regulation, (2) symbolic-technical self-regulation, and (3) comprehensive socio-technical self-regulation.

Keywords: AI governance, organization, artificial intelligence, responsible AI, guidelines, self-regulation

Auf dem Weg zur KI-Governance im DAX40: Eine Typologie organisatorischer Leitlinien für die Selbstregulierung

Zusammenfassung: In diesem Beitrag verbinden wir den Diskurs zu ethischen Herausforderungen des Einsatzes von Künstlicher Intelligenz (KI) mit Forschungsperspektiven auf eine übergeordnete KI-Governance. Obgleich diese noch nicht institutionalisiert ist, formulieren immer mehr Organisationen eigene Richtlinien für den verantwortungsvollen Einsatz von KI. Diese Selbstregulierungsansätze dienen der Orientierung für Kunden und Beschäftigte. Zugleich dienen sie der Ausrichtung organisatorischer Prozesse und Prüfmechanismen. Wenig bekannt ist, welche Unterschiede es zwischen Selbstregulierungsansätzen gibt und welche zukünftige Ausrichtung einer KI-Governance Organisationen damit antizipieren. Am Beispiel von DAX40-Unternehmen untersuchen wir, wie Organisationen KI-Richtlinien zur Selbstregulierung gestalten und welche Kriterien der KI-Ethik sie dabei berücksichtigen. Auf der Grundlage einer systematischen Suchstrategie und einer qualitativen Inhaltsanalyse identifizieren wir drei verschiedene Typen der Selbstregulierung: (1) nicht kodifizierte Selbstregulierung, (2) symbolisch-technische Selbstregulierung und (3) umfassende sozio-technische Selbstregulierung.

Stichwörter: KI-Governance, Organisation, Künstliche Intelligenz, verantwortungsvolle KI, Leitlinien, Selbstregulierung

1. Introduction

Artificial Intelligence (AI) opens new possibilities for organizations and employees in performing tasks and generating solutions. Authors emphasize positive outcomes in productivity and quality (Brynjolfsson et al., 2025). However, due to its pervasive nature (von Krogh, 2018) and the opacity of data structure and algorithms (Meske et al., 2020), interest in AI governance is growing in organizations (Hickmann & Petrin, 2021; Hickok, 2021; Stahl et al., 2022). The design of organizational AI governance is shaped by a multifaceted regulatory framework. Hard law, such as the European Artificial Intelligence Act (EU AI Act) or industry-specific regulations (e.g., the regulation on digital operational resilience in finance, DORA), defines binding legal requirements. Soft law adds further normative expectations that influence corporate behavior without formal legal force. This includes harmonized standards such as ISO 42001, which translate these requirements into concrete technical and operational guidance as well as ethical principles and guidelines like the OECD AI Principles. Finally, stakeholder pressure from investors, customers, or civil society creates additional expectations for responsible and trustworthy AI. Together, these elements form the comprehensive and increasingly legally binding regulatory environment in which organizations must position their AI governance practices (Hickmann & Petrin, 2021; Mäntymäki et al., 2022a; Agrawal & Nene, 2025; Batool et al., 2025; Maman & Feldmann, 2025).

As a subset of corporate governance, AI governance is defined as a system of rules, practices, processes, and technical measures through which organizations design, deploy, and oversee AI systems to ensure that their use aligns with legal requirements, ethical principles, external expectations, and the organization's own strategies and values. It structures how organizations manage risks, ensure responsible and trustworthy AI and integrate AI in ways that support responsible and goal-oriented organizational action (Mäntymäki et al., 2022b; Wirtz et al., 2022; Schneider et al., 2023; Schneider et al., 2024; Papagiannidis et al., 2025). The main challenge for AI governance is to translate the regulatory requirements into practical organizational processes (Mäntymäki et al., 2022a; Birkstedt et al., 2023; Agrawal & Nene, 2025; Batool et al., 2025; Papagiannidis et al., 2025).

While scholars outline conceptual frameworks for a future AI governance that address the interplay between external regulatory requirements, organizational governance measures, and internal and external effects (e.g., Mäntymäki et al., 2022a; Papagiannidis et al., 2025), there is little empirical evidence to date on how organizations actually address this topic. Distinctions about coping patterns in face of rising AI governance demands are missing. As formalized AI governance systems are not established, initial indications can be found in the increasing number of organizational AI guidelines, which are formulated as a self-regulatory mechanism to create a preliminary approach for coping with ethical challenges of AI usage while translating external requirements into internal practices (Corrêa et al., 2023; Schneider et al., 2024). A number of studies have already examined AI guidelines issued by governments, industry bodies, and non-governmental organizations (e.g., Jobin et al., 2019; Hagendorff, 2020; Corrêa et al., 2023), specifying ethical principles such as transparency, justice and fairness, non-maleficence, responsibili-

ty, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, and solidarity (Jobin et al., 2019). So far, this discourse on AI ethics remains rather detached from a governance perspective that searches for concrete corporate measures to demonstrate corporate responsibility in AI usage (Batool et al., 2025). In line with Hickman and Petrin (2021) we understand AI guidelines for self-regulation as pre-stage of an AI governance worth looking into to explore how companies prepare for future formal regulation.

The aim of this study is to examine how organizations interpret corporate responsibility for ethical challenges of AI in their self-regulation, particularly which criteria of AI ethics they address in these guidelines. We focus our analysis on the forty largest listed organizations in Germany (DAX40) since these corporations have to cope with the same governance requirements and are challenged to document their responsibility. Methodologically, the study builds on a systematic document collection in which publicly available AI-related corporate documents are identified and classified as part of the data-gathering and search strategy. Based on this classified corpus, a typology of corporate self-regulation is developed through sequential pre-analysis and qualitative content analysis and is subsequently substantiated by a contextual analysis that embeds the identified guideline types within their broader organizational and governance contexts. As a result, we explore three types: (1) non-codified self-regulation, (2) symbolic-technical self-regulation, and (3) comprehensive socio-technical self-regulation.

2. Theoretical foundations

2.1 AI guidelines as a practice of corporate self-regulation

The strategy-as-practice research community (e.g., Whittington, 2006; Jarzabkowski & Paul Spee, 2009; Seidl et al., 2024) shows that future-oriented preparation for corporate challenges is less of a formalized process and more based on forward-going and feedback processes in interaction with relevant stakeholders. This also matters for future challenges in corporate governance such as AI implementation, where organizational guidelines can be understood as a relevant practice to initiate self-regulatory mechanisms for demonstrating responsibility in implementation and usage (Camilleri, 2024; Papagiannidis et al., 2025). With their AI guidelines organizations communicate which principles they consider particularly important for the development and use of this technology. These guidelines can take various forms, such as ethical declarations, codes of conduct, white papers, or statements. These documents specify the organization's key criteria of AI ethics, but also the overall objectives, roles, authorities, and further responsibilities for the deployment and use of AI. AI guidelines can range from internal guidance to public commitments (Jobin et al., 2019; Hickok, 2021; Attard-Frost & Walters, 2022; Corrêa et al., 2023; Schneider et al., 2024; Prem, 2024; Cabiddu et al., 2025). As such, guidelines indicate how organizations interpret regulatory requirements and stakeholder demands while transferring them into concrete, organization-specific principles with impact on internal practices and stakeholder communication (Mäntymäki et al., 2022a; Schneider et al., 2024; Maman & Feldmann, 2025). AI guidelines are shaped under the influence of external requirements as well as organizational strategies and values (Papagiannidis et al., 2025).

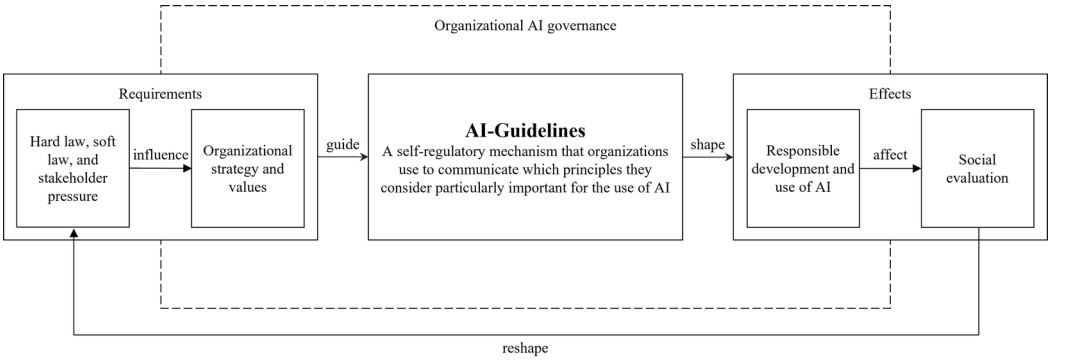


Figure 1: Requirements and effects of AI guidelines as a governance mechanism - Adapted model according to Papagiannidis et al., 2025 and Mäntymäki et al., 2022a

Following Papagiannidis et al. (2025) and Mäntymäki et al. (2022a) AI guidelines initiate an exchange basis for coping with ethical challenges of AI. It is in particular the external requirements of hard and soft law, and stakeholder pressure that form the framework for the design of AI guidelines (see Figure 1). While soft law initiatives such as the OECD AI Principles or the UNESCO Recommendations on the Ethics of AI initially dominated the regulatory discourse, the emergence of generative AI has led to the creation of legally binding regulations around the world (Alanoca et al., 2025). In Europe, the EU AI Act has developed a representative legal framework that regulates the use and development of AI in EU countries (European Parliament, 2024). In addition to this overarching legal framework, there are a number of applicable regulations that are not explicitly designed for AI systems but, for example, represent data rules such as the GDPR and therefore also apply to these systems (Viljanen & Parviainen, 2022). Taken together, these regulations provide the normative reference for organizational self-regulation that organizations in Europe have to cope with.

While AI guidelines currently remain a voluntary instrument of corporate self-regulation rather than a mandatory reporting requirement, they nevertheless represent a documented commitment to these internal and external requirements surrounding the use of AI. However, the effective translation of this commitment into organizational practice depends on complementary governance mechanisms. These include structural mechanisms that aim to define responsibilities within the organization, e.g., via an AI committee, and assign decision-making authorities. Other procedural mechanisms such as process design, performance management, and compliance monitoring serve to align decisions and actions with the organization's strategic and value-based objectives. Finally, relational mechanisms support communication and equip employees with the skills they need to use AI responsibly. This includes not only ongoing communication about AI, but also raising employee awareness of the technology, its potential and challenges, as well as the associated organizational and structural changes, and promoting their AI literacy (Schneider et al., 2023, 2024; Cabiddu et al., 2025).

2.2 Criteria of AI ethics from a socio-technical perspective

Ethical discourses specify ethical challenges and point out criteria of responsible AI implementation in organizations. Often emphasized challenges are hidden beneficence, non-maleficence, autonomy, justice, and explicability (e.g., Floridi et al., 2018; Prem 2023; Herrmann & Pfeiffer, 2023).

At the same time scholars go beyond a list of criteria and elaborate on context-specific consideration of responsible AI application (Widder & Nafus, 2023; Herrmann & Pfeiffer, 2023). This is routed in socio-technical systems thinking (Nitsch et al., 2024) reflecting the different ways of enacting a technology under organization-specific conditions (Orlikowski & Scott, 2008; Leonardi & Treem, 2020). Defining responsible AI may thus include technological characteristics as an input factor for a corporate service, job design as a throughput factor of what happens to employees interacting with the technology or effected by the AI system, and consequences for customers or clients as a typical outcome factor (Parker & Grote, 2022; Bankins & Formosa, 2023; Berretta et al., 2023). As different discourses and disciplines highlight distinct criteria of responsible AI, Wilkens et al. (2023) provide a comprehensive overview describing which criteria are associated by which target group. Trustworthiness, privacy and ethics as well as explainability are criteria facing challenges in technology development. These challenges are frequently discussed in computer and information science for AI developers. The criteria of job loss prevention, physical and mental health as well as human agency and augmentation are related to the challenges AI enhances in employee development and job crafting. They are

Technology development	**Explainability**	Transparent data usage and interpretation to improve technology adoption and to provide helpful information to users (e.g., remaining error probabilities)
	Trustworthiness, privacy & ethics	Unbiased data structure and ethical concerns in data collection and usage, with the aim of operating AI reliably and ethically without discrimination
Organizational development	**Accountability & safety culture**	Establishment of systems and organizational routines (e.g., process descriptions or checklists) to ensure reliability and to promote responsibility at system level
	Compensation of weaknesses in the system	Deficit-oriented view to compensate for human fatigue, unstable concentration or cognitive limitations in sensory discrimination
	Knowledge utilization from the user domain	Close integration of the user domain in software development
Employee development	**Augmentation & human agency**	Technology design for an enhanced use by employees who experience empowerment and professionalization through the human-AI interaction
	Physical & mental health	Protecting employees from negative influences such as heavy loads, chemical substances, or stressful interactions
	Job loss prevention	Prevention from negative consequences of new technologies on employment

Table 1: Criteria of responsible AI explored from a transdisciplinary literature review (Wilkens et al., 2023)

often discussed in psychology, sociology and HRM facing the responsibility of supervisors and employee representatives. In addition to that, the criteria accountability and safety culture, compensation of weaknesses in the system, and knowledge utilization from the user domain face challenges of AI concerning organizational design as discussed in engineering studies and organizational science (see Table 1).

In summary, an evaluation of AI guidelines for self-regulation as elaborated in this paper should consider what hard and soft law factors are taken into consideration, what mechanisms of self-regulation in terms of authorities and procedures are implemented, and what concrete (context-specific) criteria of AI ethics are specified.

3. Methodology

3.1 Sampling approach, research design, and data collection

To examine how organizations practice self-regulation through AI guidelines, we systematically evaluated publicly available AI guidelines from the 40 organizations listed on the German share index (DAX40). These organizations have the highest standards in terms of reporting and general corporate governance. Therefore, they should also be the first to report according to a defined AI governance standard. With this selection approach, we were able to ensure a uniform legal framework for all companies (e.g., EU AI Act and General Data Protection Regulation, GDPR), while also including different industries, comparable company sizes, and a common country culture.

We conducted a systematic search strategy (Snyder, 2019) and followed common standards for the structured selection of written material (Page et al., 2021). We searched for information on corporate websites, public archives, PDF documents, financial and annual reports, ethical guidelines, as well as external reports or articles, as AI guidelines are not necessarily explicitly labeled as such. In addition to this manual search, we use Boolean operators to combine the keywords *artificial intelligence, AI, machine learning, ML*, or *intelligent technology* with the keywords *guideline, declaration, standards, digital strategy, framework, whitepaper,* or *regulation* in Google's search engine and in the search functions of the corporate websites. We excluded documents that did not specifically relate to the use of AI within the organization or that merely provided general examples of AI use without addressing the underlying intentions or governance practices for its use within the organization, as well as documents written in languages other than English or German. The systematic data collection took place from October to November in 2023, followed by further refinement from October to November in 2024 to include potential reactions of the entry into force of the EU AI Act in August 2024. Since the database for the two collection dates did not differ significantly, no differentiation is made between the collection dates in the analysis.

Following the systematic data collection, 95 documents were identified for in-depth analysis and then classified according to document type and thematic focus. Documents that explicitly deal with ethical principles for the development and use of AI and are presented as stand-alone or clearly defined policy documents were categorized as AI guidelines. These primarily include dedicated AI codes, codes of conduct, and white papers. In addition, other corporate publications such as annual reports or sustainability reports, as well as information on the company website, were classified as supplementary contextual documents if they contained references to AI-related activities, structures, or responsibili-

ties. This classification served to structure the heterogeneous document corpus and create a transparent database for subsequent analysis procedures. All analyzed information and documents are stored as PDF-files in a digital repository and are available on request from the authors (see Figure 2).

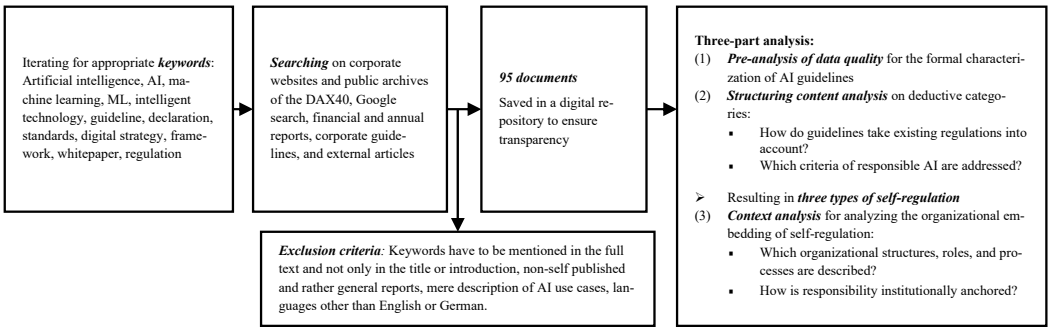


Figure 2: Flow chart of the search strategy and the evaluation process - Flow chart according to Page et al. (2021)

3.2 Data coding and data analysis

The data analysis followed a semi-systematic three-part process combining qualitative and quantitative indicators (Snyder, 2019). As a pre-analysis, we distinguished the quality of the data source, as it is not identical for all DAX40 companies, but indirectly indicates to what extent and in what depth companies show awareness and give attention to ethical challenges of AI. The varying quality of the available data is therefore provided as supplementary information and is based exclusively on documents classified as dedicated AI guidelines. Based on this, we assessed in the main evaluation the extent to which AI guidelines take existing regulations into account. We coded what regulations are explicitly mentioned in the guidelines, specifically referring to the EU AI Act, national laws, and industry-specific regulations that are referenced in the context of AI use, such as DORA. Moreover, we examined which criteria of responsible AI are addressed in the guidelines. We referred to the catalogue of criteria and their underlying definitions as outlined in Table 1. The coding refers to the meaning in corporate documents more than to explicit wording as this is not always identical to the scientific outlines. Building on this analytical procedure, the variations in how organizations address the criteria reveal distinct patterns of self-regulation. These patterns consolidate into three overarching types of self-regulatory approaches, which differ in external explicitness of their guidelines and considered criteria of AI ethics.

In the third part of the data analysis, we drew on supplementary documents to contextualize how self-regulatory efforts are implemented beyond formalized guideline mechanisms. We focused on the organizational structures, roles, and processes described, and on how responsibility is institutionally anchored. (Schneider et al., 2023, 2024). Based on a structuring content analysis (Mayring & Fenzl, 2014; Mayring, 2019), the coding scheme is used independently by two coders in a deductive manner, considering an intercoder-reliability (Lombard et al., 2022).

4. Findings

The following subsections outline the three identified types of AI self-regulation in sequential order (see Table 2). *Type 1 – Non-codified self-regulation* comprises approaches without publicly accessible AI guidelines, but which do provide information on governance measures. *Type 2 – Symbolic technical self-regulation* includes guidelines that translate selected principles into technical artifacts or processes, but are often only partial or illustrative in nature. *Type 3 – Comprehensive socio-technical self-regulation* encompasses guidelines that pursue a distinct socio-technical perspective by linking ethical principles with concrete organizational and technical measures, thereby demonstrating how these principles are implemented in daily operations. These characterizations provide an initial overview, which will be further detailed and empirically substantiated in the course of this paper.

	DAX40 Organizations	Segment	Pre-analysis of data quality: Formal Preparation	Pre-analysis of data quality: Scope & Depth	Content Analysis: Reference of existing regulations relating to the use of AI	Explain-ability*ted	Trustworthiness, privacy & ethics*ted	Accountability & safety culture*od	Compensation of weaknesses in the system*od	Knowledge utilization from the user domain*od	Augmentation & human agency*ed	Physical & mental health*ed	Job loss prevention*ed
Type 3- Comprehensive socio-technical self-regulation	Allianz	Finance	Guidelines on homepage	Comprehensive	Orientation toward existing regulations	x	x	x		x	x		
	Bayer	Chemicals, Pharmaceuticals, Medical Technology	Code of Conduct		Orientation toward existing regulations	x	x	x		x	x	x	
	BMW	Mechanical engineering, transportation, logistics	AI Ethics Code		Orientation toward existing regulations, explicit reference to EU AI Act	x	x	x			x	x	
	Commerzbank	Finance	AI Principles		Orientation toward existing regulations, explicit reference to EU AI Act	x	x	x			x	x	x
	Continental	Mechanical engineering, transportation, logistics	Guidelines on homepage		Orientation toward existing regulations, explicit reference to *Ethics Guidelines for Trustworthy AI* (AI HLEG) and *Ethically Aligned Design* (IEEE)	x	x	x			x	x	
	Deutsche Telekom	Electronics, hardware, software	Digital Ethics Guidelines on AI		Orientation toward existing regulations, explicit reference to EU AI Act	x	x	x	x	x	x	x	x
	Infineon	Electronics, hardware, software	AI Manifest		Orientation toward existing regulations	x	x	x	x		x	x	
	Mercedes-Benz Group	Mechanical engineering, transportation, logistics	Guidelines on homepage		Orientation toward existing regulations	x	x	x			x		x
	Merck	Chemicals, Pharmaceuticals, Medical Technology	Code of Digital Ethics		Orientation toward existing regulations	x	x	x		x	x	x	
	SAP	Electronics, hardware, software	Global AI Ethics Policy		Orientation toward existing regulations	x	x	x			x		
	Volkswagen	Mechanical engineering, transportation, logistics	Ethical Principles for AI		Orientation toward existing regulations	x	x	x			x		x

			Pre-analysis of data quality		Content Analysis									
	DAX40 Organizations	Segment	Formal Preparation	Scope & Depth	Reference of existing regulations relating to the use of AI	Explain-ability*ed	Trustworthiness, privacy & ethics*ed	Accountability & safety culture*ed	Compensation of weaknesses in the system*od	Knowledge utilization from the user domain*od	Augmentation & human agency*ed	Physical & mental health*ed	Job loss prevention*ed
Type 2 – Symbolic-technical self-regulation	Airbus	Mechanical engineering, transportation, logistics	Guidelines on homepage	Superficial	Orientation toward existing regulations, explicit reference to *Ethics Guidelines for Trustworthy AI* (AI HLEG) and *Trustworty AI building blocks* (EASA)	(x)	(x)	(x)			(x)	(x)	(x)
	BASF	Chemicals, Pharmaceuticals, Medical Technology	Code of Conduct		Orientation toward existing regulations	(x)	(x)					(x)	
	Daimler Truck	Mechanical engineering, transportation, logistics	Code of Conduct		Orientation toward existing regulations	(x)	(x)	(x)					(x)
	Deutsche Bank	Finance	Guidelines on homepage		Orientation toward existing regulations, explicit reference to *Ethics Guidelines for Trustworthy AI* (AI HLEG)	(x)	(x)	(x)			(x)		
	E.ON	Energy and raw materials	White Paper		Orientation toward existing regulations, explicit reference to *GDPR*	(x)	(x)		(x)				
	Henkel	Chemicals, Pharmaceuticals, Medical Technology	Code of Conduct		Orientation toward existing regulations, explicit reference to *GDPR*	(x)	(x)	(x)					
	Siemens	Electronics, hardware, software	Sustainability Report		Orientation toward existing regulations, explicit reference to *EU AI Act* and *GDPR*	(x)	(x)	(x)			(x)		
	Siemens Healthineers	Chemicals, Pharmaceuticals, Medical Technology	Sustainability Report		Orientation toward existing regulations, explicitly reference to *EU AI Act* and *GDPR*	(x)	(x)	(x)			(x)		

	DAX40 Organizations	Segment	Pre-analysis of data quality: Formal Preparation	Pre-analysis of data quality: Scope & Depth	Content Analysis: Reference of existing regulations relating to the use of AI	Content Analysis: Explain-ability*td	Content Analysis: Trustworthiness, privacy & ethics*td	Content Analysis: Accountability & safety culture*od	Content Analysis: Compensation of weaknesses in the system*od	Content Analysis: Knowledge utilization from the user domain*od	Content Analysis: Augmentation & human agency*ed	Content Analysis: Physical & mental health*ed	Content Analysis: Job loss prevention*ed
Type 1 – Non-codified self-regulation	Beiersdorf	Chemicals, Pharmaceuticals, Medical Technology	no formal preparation	no formal preparation	Orientation toward existing regulations								
	Covestro	Chemicals, Pharmaceuticals, Medical Technology			Orientation toward existing regulations								
	MTU Aero Engines	Mechanical engineering, transportation, logistics			Orientation toward existing regulations								
	Münchner Rück	Finance			Orientation toward existing regulations, explicitly reference to *Ethics Guidelines for Trustworthy AI* (AI HLEG)								
	Porsche	Mechanical engineering, transportation, logistics			Orientation toward existing regulations								
	RWE	Energy and raw materials			Orientation toward existing regulations								
	adidas	Retail and Consumer Goods			No explicit reference to AI related regulation								
	Brenntag	Chemicals, Pharmaceuticals, Medical Technology			No explicit reference to AI related regulation								
	Deutsche Börse	Finance			Orientation toward existing regulations and positioning to the European Comission's *White Paper on AI – A European Approach*								
	Deutsche Post	Mechanical engineering, transportation, logistics			Orientation toward existing regulations								
	Fresenius	Chemicals, Pharmaceuticals, Medical Technology			No explicit reference to AI related regulation								
	Hannover Rück	Finance			Orientation toward existing regulations, explicitly reference to *EU AI Act*								

	DAX40 Organizations	Segment	Pre-analysis of data quality: Formal Preparation	Pre-analysis of data quality: Scope & Depth	Content Analysis: Reference of existing regulations relating to the use of AI	Content Analysis: Explain-ability*td	Content Analysis: Trustworthiness, privacy & ethics*td	Content Analysis: Accountability & safety culture*od	Content Analysis: Compensation of weaknesses in the system*od	Content Analysis: Knowledge utilization from the user domain*od	Content Analysis: Augmentation & human agency*ed	Content Analysis: Physical & mental health*ed	Content Analysis: Job loss prevention*ed
Type 1 – Non-codified self-regulation	Heidelberg Materials	Others			Orientation toward existing regulations								
	Porsche Automobil Holding	Mechanical engineering, transportation, logistics			No explicit reference to AI related regulation								
	Qiagen	Chemicals, Pharmaceuticals, Medical Technology			No explicit reference to AI related regulation								
	Rheinmetall	Mechanical engineering, transportation, logistics			No explicit reference to AI related regulation								
	Sartorius	Chemicals, Pharmaceuticals, Medical Technology			No explicit reference to AI related regulation								
	Siemens Energy	Energy and raw materials			No explicit reference to AI related regulation								
	Symrise	Chemicals, Pharmaceuticals, Medical Technology			No explicit reference to AI related regulation								
	Vonovia	Real estate			No explicit reference to AI related regulation								
	Zalando	Retail and Consumer Goods			No explicit reference to AI related regulation								

Table 2: Content-related emphasis of self-regulation according to the types of AI guidelines of the DAX40 – Criteria of Wilkens et al. (2023): *td = technology development; *od = organizational development; *ed = employee development. Fulfilment of criteria: empty cell = no explicit reference; (x) = superficially addressed; x = comprehensively addressed

Type 1 – Non-codified self-regulation

Pre-analysis of data quality and content analysis: No formal preparation of defined guidelines or criteria for AI use could be identified from the data material for companies of this type (n = 21). For this reason, the database does not allow for a more in-depth content analysis. It is noteworthy that in individual cases (n = 6), organizations indicate that internal guidelines for the use of AI have been developed. For example, Beiersdorf writes in its 2023 annual report that the company "has published binding legal guidelines within a short period of time that all Beiersdorf employees must observe when using [generative AI] applications" (Beiersdorf, Annual Report 2023, p. 172).

Contextual analysis: Apart from the lack of specific guidelines, there are references in individual cases to existing AI-related regulations and governance measures. Ten of the 21 companies indicate that they take existing regulations into account, with three cases explicitly referring to specific regulations. For example, Hannover Rück emphasizes in its annual report that the company intends to adapt its existing guidelines on the use of AI to the legal requirements of the EU AI Act and is preparing for possible guidelines on AI governance and risk management. Similarly, Münchener Rück refers to efforts at the European level and states in its 2023 annual report that it has developed its own strategy in line with the European Commission's guidelines. In a position paper, Deutsche Börse Group states: "[Most] activities/services provided by AI applications in the financial sector would be regulated by existing rules and laws" (Deutsche Börse Group, Comments on EC's communication, p. 2), which include, among others, the Markets in Financial Instruments Directive II (MiFID II) and the Markets in Financial Instruments Regulation (MiFIR) framework.

In addition, several measures for governing AI are mentioned ranging from individual measures to comprehensive control mechanisms. These include cooperation and strategic partnerships with external partners such as IBM (Deutsche Post) or DFKI (Sartorius) or references to formal governance structures, such as the establishment of a research lab at Sartorius. In the spirit of participation and co-determination, the Heidelberg Materials works council has also taken a position on the use of AI and its possible consequences within the company. Covestro, on the other hand, refers to employee training on the opportunities, risks, and reflective use of AI. In two selected cases, a combination of several measures became apparent, such as MTU Aero Engines, which refers to its collaboration with applied AI to develop not only an AI strategy, but also a competence center that promotes implementation guidelines, training concepts, and measures to raise awareness of AI within the company. RWE's measures are similarly comprehensive and comprise an AI research laboratory, an AI control framework to guide and support future AI projects, and the involvement of all necessary stakeholders and employee representatives. This reflects the efforts of organizations to address the internal use of AI.

Due to the lack of dedicated guidelines and isolated references to governance measures, we refer to this type as *non-codified self-regulation*. The crucial point is that self-regulation mechanisms evolve, while the content of what is regulated in particular remains unspecified for externals.

Type 2 - Symbolic-technical self-regulation

Pre-analysis of data quality: Organizations of this type (n = 8) report in greater detail on their use of AI and provide specific examples of its application. They choose different ways of presenting this information, ranging from general references to guidelines on their homepages to official documents such as codes of conduct or annual reports, in which they address the use of AI in the company in specific paragraphs and refer to superficial criteria.

Content analysis: With regard to the consideration of existing legal and regulatory frameworks in relation to AI, all organizations refer either implicitly to alignment with existing regulations or explicitly to regulatory efforts at the European level, such as the Ethics Guidelines for Trustworthy AI from the High-Level Expert Group on AI (AI HLEG) or the EU AI Act. This is illustrated by the example of Siemens, which applies "generally accepted and trustworthy AI frameworks to AI tools [in anticipation of the] upcoming AI regulation in the EU" (Siemens, Sustainability Report 2023, p. 35). Airbus stands out in particular by referring to the industry-specific regulations of the European Aviation Safety Agency (EASA), which has developed trustworthy AI building blocks for the aviation industry based on the Ethics Guidelines for Trustworthy AI, which are relevant for the development of aircraft.

With regard to the anchoring of ethical principles, the data reveals loosely listed criteria for the use of AI, as well as descriptions formulated by organizations for dealing with ethical challenges. The data shows that all organizations refer to the criteria of technology development in terms of *explainability* and *trustworthiness, privacy & ethics*, followed by the criteria of *accountability & safety culture*, and *augmentation & human agency*, in their statements. Other criteria are given less consideration, and the criterion of *knowledge utilization from the user domain* is not considered at all. In the case of Airbus, for example, it was possible to assign criteria using defined principles such as "accountability and transparency" and "safety first," which refer to the criteria of *accountability & safety culture*. In cases where there were no explicit criteria, but only descriptions, it was possible to classify them based on the indirect meaning of the statement. This becomes clear in the case of BASF, where the statements in the code of conduct to the effect that the company uses technologies inclusively as part of its digital responsibility and seeks to avoid reinforcement effects and unfair distortions could be linked to the criterion of *trustworthy AI*.

Contextual analysis: Apart from ethical considerations, the documents also refer to measures like cooperation and strategic partnerships (e.g., BASF) and include participation and co-determination initiatives like the involvement of employee representatives and the development of a group agreement to address AI-induced changes (e.g., Daimler Truck). In addition, measures to prepare and train employees are mentioned, as well as formal governance structures and oversight initiatives, through Deutsche Bank's Artificial Intelligence Oversight Forum (AIOF), to "ensure appropriate monitoring and risk assessment of AI solutions and their alignment with the bank's strategic goals" (Deutsche Bank, non-financial report 2023, p. 141).

Since the individual criteria are not explained in detail and mainly refer to technological aspects, with only partial reference to supplementary governance measures, we refer to this type as *symbolic-technical self-regulation*.

Type 3 - Comprehensive socio-technical self-regulation

Pre-analysis: Organizations of this type (n = 11) publish comprehensive guidelines on their websites, in documents such as a code of conduct or in explicit AI documents (e.g., AI manifesto, AI code, or similar), in which they define principles for the responsible use of AI and supplement these with further explanations on how to translate them into operational practice.

Content analysis: With regard to the consideration of existing legal and regulatory frameworks in relation to AI, companies are adopting approaches that vary in scope. Bayer, for example, points out that AI is used within the applicable legal and ethical framework. Continental devotes a separate section to this topic in its Code of Ethics for AI and addresses laws, regulations, as well as Continental's own rules, standards, and instructions. Deutsche Telekom, in turn, mentions in its annual report that it is preparing for the implementation of the EU AI Act, and Commerzbank is aligning its efforts with the EU AI Act and comments on regulatory developments at the European level in a separate paper.

Concerning the anchoring of ethical principles, all corporations emphasize criteria of technology development, as well as *accountability and safety culture* and *augmentation and human agency* referring to organizational and personnel development. While the criterion *physical and mental health* is still considered comparatively often, the remaining criteria are considered less often. The organizations do not only specify principles, but also explain how they are implemented in practice and, in some cases, who is responsible for their further development and compliance within the company. The Mercedes-Benz Group, for example, addresses the criterion of explainability by emphasizing its commitment to maintaining a high level of transparency in the use of AI in order to promote trust in AI systems and gives further explanations of how this is achieved.

Contextual analysis: With regard to governance initiatives, type 3 companies include more references and are increasingly focusing on establishing formal accountability structures and oversight, as well as organizational and normative anchoring, which aim to implement the defined guidelines and embedded principles for the use of AI in the culture of the organization. These include dedicated organizational units such as a data advisory board (Allianz), various working groups and expert groups (Deutsche Telekom), and committees, panels and offices with different areas of responsibility (SAP). In addition to this, references can be found to the founding of competence centers (e.g. BMW). Particularly noteworthy are the approaches taken by Continental, Deutsche Telekom, and SAP. Continental defines so-called “AI owners”, who are responsible for implementing the guidelines and ensuring compliance withing the company. Deutsche Telekom and SAP complement their guidelines with comprehensive measures, such as interdisciplinary working groups, expert panels on specific topics such as ChatGPT, prompt-a-thons to promote a culture of enablement within the company, assessments to ensure the implementation of AI requirements, and many more.

Since these companies take all socio-technical dimensions into account in their guidelines and supplement them with organizational initiatives for the responsible use of AI, we refer to them as c*omprehensive socio-technical self-regulation*.

5. Discussion and limitations

The debate on AI ethics is not uniformly aligned with corporate governance. In our analysis, we evaluate the self-regulation documents of DAX40 companies to gain deeper insight into how these companies are preparing for future governance requirements and to what extent they refer to criteria that are considered in the debate on AI ethics.

The analysis shows that DAX40 companies have a common ground for designing and applying AI guidelines in terms of compliance with existing regulations, the establishment of predominantly technical ethical principles, and supplementary governance initiatives. Differences primarily exist in the formal preparation of the guidelines and the criteria that go beyond technical aspects, from which three types of self-regulation could be derived. Type 1 is characterized by the absence of a dedicated and publicly available AI guideline, with ethical challenges related to AI addressed only indirectly or in a fragmented manner across other corporate documents. Type 2 companies articulate explicit ethical principles within standalone AI guidelines, yet limit their scope to a selective set of ethical challenges. In contrast, type 3 reflects a comprehensive socio-technical approach to AI self-regulation, engaging with a broad range of ethical considerations discussed in the AI ethics literature and systematically linking them to organizational practices and governance structures. This suggests that type 3 companies are highly aware of AI challenges in their operational business.

The differences identified between the types cannot be attributed entirely to industry affiliation or other similarities in contextual factors. However, it is interesting to note that companies in the electronics, hardware, and software sectors are particularly strongly represented in type 3. The core business of these companies revolves around data, which means that AI is more than just a tool for a support process. It is often part of the solution offered. Other companies represented in type 3 have a long tradition of co-determination by works councils, as it is known from the automotive industry. Further interpretation of the antecedents of different types requires further future research with more in-depth analysis of the corporations.

While companies are currently fundamentally free to design their own AI self-regulation, DAX40 companies are used to foster transparency for governance reasons due to their size, regulatory interdependence, and economic and social significance in other governance fields. Thus, they are expected to go beyond minimum standards of co-determination and prepare for future governance requirements. As long as there is no overarching legally obliged standard in reporting, self-regulation might even help to differentiate while demonstrating early adoption and high responsibility for AI usage as a strategic factor.

Against the backdrop of corporate governance research in general (Di Vito & Trottier, 2022; Mueller, 2022) and AI governance research in particular (Birkstedt et al., 2023; Batool et al, 2025), our findings underscore that the self-regulatory mechanisms of checks and balances for addressing ethical challenges are at the heart of new governance issues. It is AI governance as an incremental feed-forward practice that sets in motion a gradual process of further development. Even if companies remain rather vague about the importance of AI ethics, they are developing control bodies and reporting mechanisms to demonstrate their responsibility. In line with Hickman and Petrin (2021), we therefore conclude that self-regulatory mechanisms should be considered and recognized as a starting point for AI governance, even if they cannot replace a comprehensive AI governance structure.

By analyzing AI guidelines for self-regulation among DAX40 companies and developing a typology, we build on former scholars' work on AI governance (e.g., Mäntymäki et al., 2022a; Schneider et al., 2024; Papagiannidis et al., 2025) by making previously undefined forms of corporate self-regulation visible with the help of a comparative content analysis. Our typology not only reflects differences identified in earlier work (e.g., Jobin et al., 2019; Hagendorff, 2020), but also clarifies their origin. Differences result from a varying scope of ethical criteria addressed and the already implemented governance mechanisms to monitor effects and operate control. In doing so, we also build an empirical bridge between established governance models and corporate self-regulation activities indicating contextualized measures that can be considered as pre-stage practices towards an AI governance.

The analysis can be considered as a starting point for further research. We would first like to encourage empirical validation to investigate whether advancements in self-regulation or exceeding minimum standards offer advantages either in organizational internal technology acceptance or as external competitive advantages in the different markets of the DAX40. We assume that there are differences between markets, whether high standards in AI ethics matter or not. Thus, it will be interesting to find out for which industries and under what conditions high AI governance standards may have competitive impact. The contribution of this paper is to make empirically validated distinctions of the independent variable while specifying different types of coping with challenges of AI ethics.

With respect to regulatory research, it remains an open and important question to what extent socio-technical criteria will be considered for future legal frameworks, and which initiatives may foster a more nuanced socio-technical understanding of AI ethics in regulatory communities.

The typology developed in this paper helps politicians, companies, and scientists alike. For politicians, the different types indicate whether and to what degree companies cope with ethical challenges in a manner that demonstrates high self-obligation. Only about a quarter of DAX40 corporations anticipate risks of broader scope, while most corporations neglect in particular social challenges when implementing AI. This is an important insight for the development of legal obligations. In addition, the typology provides information for managers responsible for corporate decision-making on which criteria of AI ethics and mechanisms for checks and balances can be activated to demonstrate a high level of risk awareness. The study reveals that AI self-regulation can be considered strategically when going beyond minimum requirements, especially as long as general standards for all corporations are not legally binding. Moreover, the typology can serve as a guide for step-by-step implementation of an AI governance structure and highlight, which measures (e.g., ethics council, risk management, etc.) are crucial for documenting advancements. At the same time, the study underscores the importance of actively anticipating regulatory developments, choosing a clear governance logic, operationalizing guidelines, and aligning the intent of AI use with the guidelines and governance measures.

Like any other study, this study is not without limitations. First, the generalizability of the results is limited. The analysis of DAX40 corporations was a conscious sampling decision to focus on those corporations where one can expect the most advanced preparations for AI governance, in order to analyze not the mean but the leading edge of corporations. Since only a minority of DAX40 corporations have developed advanced self-regulatory

approaches, it can be assumed that such standards are even less established among other companies. Second, the focus on publicly available information and documents carries the risk of incompleteness. We have taken this risk into account by also highlighting in type 1 that individual companies have guidelines that are not directly accessible, but whose existence can be verified through references and accompanying governance measures. Third, the time perspective is a crucial variable, especially in this dynamic subject area. Policies are likely to evolve dynamically in line with technical and regulatory developments, meaning that at the time of analysis, we were only able to gain a snapshot insight into the state of self-regulation. However, we have attempted to counteract this by translating the individual results of the companies into an overarching typology that represents a more time-independent systematization and sharpens the theoretical understanding of the design and use of AI policies in organizations.

6. Conclusion and outlook

Organizational AI guidelines for self-regulation are just one component of a complex regulatory approach that includes soft law components, legally binding regulations, and practical implementation to promote the ethical use and development of AI (Batool et al., 2025; Maman & Feldman, 2025). This study shows that AI guidelines and complementary governance measures represent an approach to corporate self-regulation that can both promote the responsible use of AI in organizations and contribute to the overarching social discourse and regulatory developments. It becomes evident that it is not only the mere existence of guidelines that matters, but also their design, referenced ethical principles, and supplementation with additional governance measures.

AI is increasingly finding its way into corporate reporting and is already being treated as a central component of future corporate social responsibility reporting (e.g., Camilleri, 2024). With comprehensive socio-technical approaches to self-regulation, organizations are taking a significant step toward comprehensive AI governance structures.

Ultimately, however, the potential and effectiveness of these approaches can only be verified by looking at internal governance processes. We therefore encourage future research to empirically validate whether advancements in self-regulation provide any organizational internal or market-based advantage.

References

Agarwal, A., & Nene, M. J. (2025). A five-layer framework for AI governance: Integrating regulation, standards, and certification. *Transforming Government: People, Process and Policy*, *19*(3), 535–555. https://doi.org/10.1108/TG-03-2025-0065

Alanoca, S., Gur-Arieh, S., Zick, T., & Klyman, K. (2025). Comparing Apples to Oranges: A Taxonomy for Navigating the Global Landscape of AI Regulation. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 914–937. https://doi.org/10.1145/3715275.3732059

Attard-Frost, B., & Walters, D. R. (o. J.). The Ethics of AI Business Practices: A Review of 47 AI Ethics Guidelines. *AI Ethics* 3, 389–406 (2023). https://doi.org/10.1007/s43681-022-00156-6

Bankins, S., & Formosa, P. (2023). The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work. *Journal of Business Ethics*, *185*(4), 725–740. https://doi.org/10.1007/s10551-023-05339-7

Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00653-w

Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., & Kluge, A. (2023). Defining human-AI teaming the human-centered way: A scoping review and network analysis. *Frontiers in Artificial Intelligence*, *6*, 1250725. https://doi.org/10.3389/frai.2023.1250725

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, *33*(7), 133–167. https://doi.org/10.1108/INTR-01-2022-0042

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work. *The Quarterly Journal of Economics*, *140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

Cabiddu, F., Lauro, S. D., Samaan, D., & Tursunbayeva, A. (o. J.). Governing AI in the World of Work: An International Review of 245 Ethics Guidelines. *Available at SSRN*. http://dx.doi.org/10.2139/ssrn.5397353

Camilleri, M. A. (2024). Artificial intelligence governance: Ethical considerations and implications for social responsibility. *Expert Systems*, *41*(7), Article 7. https://doi.org/10.1111/exsy.13406

Corrêa, N. K., Galvão, C., Santos, J. W., Del Pino, C., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & De Oliveira, N. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, *4*(10), 100857. https://doi.org/10.1016/j.patter.2023.100857

Di Vito, J., & Trottier, K. (2022). A Literature Review on Corporate Governance Mechanisms: Past, Present, and Future*. *Accounting Perspectives*, *21*(2), 207–235. https://doi.org/10.1111/1911-3838.12279

European Partliament (2024). Artificial Intelligence Act (Regulation (EU), 2024/1689)

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), Article 1. https://doi.org/10.1007/s11023-020-09517-8

Herrmann, T., & Pfeiffer, S. (2023). Keeping the organization in the loop: A socio-technical extension of human-centered artificial intelligence. *AI & SOCIETY*, *38*(4), 1523–1542. https://doi.org/10.1007/s00146-022-01391-5

Hickman, E., & Petrin, M. (2021). Trustworthy AI and Corporate Governance: The EU's Ethics Guidelines for Trustworthy Artificial Intelligence from a Company Law Perspective. *European Business Organization Law Review*, *22*(4), Article 4. https://doi.org/10.1007/s40804-021-00224-0

Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI and Ethics*, *1*(1), Article 1. https://doi.org/10.1007/s43681-020-00008-1

Jarzabkowski, P., & Paul Spee, A. (2009). Strategy-as-practice: A review and future directions for the field. *International Journal of Management Reviews*, *11*(1), 69–95. https://doi.org/10.1111/j.1468-2370.2008.00250.x

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), Article 9. https://doi.org/10.1038/s42256-019-0088-2

Leonardi, P. M., & Treem, J. W. (2020). Behavioral Visibility: A new paradigm for organization studies in the age of digitization, digitalization, and datafication. *Organization Studies*, *41*(12), 1601–1625. https://doi.org/10.1177/0170840620970728

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, *28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Maman, L., & Feldman, Y. (2025). Compliance and Effectiveness of Industry Self-Regulation: A Systematic Literature Review. *Available at SSRN*. http://dx.doi.org/10.2139/ssrn.5233166

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022a). *Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance* (No. arXiv:2206.00335; Nummer arXiv:2206.00335). arXiv. https://doi.org/10.48550/arXiv.2206.00335

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022b). Defining organizational AI governance. *AI and Ethics*, *2*(4), Article 4. https://doi.org/10.1007/s43681-022-00143-x

Mayring, P., & Fenzl, T. (2014). Qualitative Inhaltsanalyse. In N. Baur & J. Blasius (Eds.), Handbuch Methoden der empirischen Sozialforschung (pp. 543–556). Springer. https://doi.org/10.1007/978-3-531-18939-0_38

Mayring, P. (2019). Qualitative Inhaltsanalyse – Abgrenzungen, Spielarten, Weiterentwicklungen. *Forum Qualitative Social Research*, 20(3), 15. https://doi.org/10.17169/fqs-20.3.3343

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2022). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, *39*(1), 53–63. https://doi.org/10.1080/10580530.2020.1849465

Mueller, B. (2022). Corporate Digital Responsibility. *Business & Information Systems Engineering*, *64*(5), 689–700. https://doi.org/10.1007/s12599-022-00760-0

Nitsch, V., Rick, V., Kluge, A., & Wilkens, U. (2024). Human-centered approaches to AI-assisted work: The future of work? *Zeitschrift Für Arbeitswissenschaft*, *78*(3), 261–267. https://doi.org/10.1007/s41449-024-00437-2

Orlikowski, W. J., & Scott, S. V. (2008). 10 Sociomateriality: Challenging the Separation of Technology, Work and Organization. *Academy of Management Annals*, *2*(1), 433–474. https://doi.org/10.5465/19416520802211644

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, *34*(2), 101885. https://doi.org/10.1016/j.jsis.2024.101885

Parker, S. K., & Grote, G. (2022). Automation, Algorithms, and Beyond: Why Work Design Matters More Than Ever in a Digital World. *Applied Psychology*, *71*(4), 1171–1204. https://doi.org/10.1111/apps.12241

Prem, E. (2023). From ethical AI frameworks to tools: A review of approaches. *AI and Ethics*, *3*(3), 699–716. https://doi.org/10.1007/s43681-023-00258-9

Prem, E. (2024). Approaches to Ethical AI. In H. Werthner, C. Ghezzi, J. Kramer, J. Nida-Rümelin, B. Nuseibeh, E. Prem, & A. Stanger (Hrsg.), *Introduction to Digital Humanism* (S. 225–239). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-45304-5_15

Schneider, J., Abraham, R., Meske, C., & Vom Brocke, J. (2023). Artificial Intelligence Governance For Businesses. *Information Systems Management*, *40*(3), Article 3. https://doi.org/10.1080/10580530.2022.2085825

Schneider, J., Abraham, R., & Meske, C. (2024). Governance of generative artificial intelligence for companies. arXiv preprint arXiv:2403.08802. https://doi.org/10.48550/arXiv.2403.08802

Seidl, D., Ma, S., & Splitter, V. (2024). What makes activities strategic: Toward a new framework for strategy-as-practice research. *Strategic Management Journal*, *45*(12), 2395–2419. https://doi.org/10.1002/smj.3668

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K., & Jiya, T. (2022). Organisational responses to the ethical issues of artificial intelligence. *AI & SOCIETY*, *37*(1), Article 1. https://doi.org/10.1007/s00146-021-01148-6

Viljanen, M., & Parviainen, H. (2022). AI Applications and Regulation: Mapping the Regulatory Strata. *Frontiers in Computer Science*, *3*, 779957. https://doi.org/10.3389/fcomp.2021.779957

Von Krogh, G. (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries*, 4(4), 404–409. https://doi.org/10.5465/amd.2018.0084

Whittington, R. (2006). Completing the Practice Turn in Strategy Research. *Organization Studies*, *27*(5), 613–634. https://doi.org/10.1177/0170840606064101

Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, *10*(1), Article 1. https://doi.org/10.1177/20539517231177620

Wilkens, U., Lupp, D., & Langholf, V. (2023). Configurations of human-centered AI at work: Seven actor-structure engagements in organizations. *Frontiers in Artificial Intelligence*, *6*, 1272159. https://doi.org/10.3389/frai.2023.1272159

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly*, *39*(4), 101685. https://doi.org/10.1016/j.giq.2022.101685

Niklas Obermann, M. Sc., is Ph.D. Candidate and Research Assistant at Ruhr University Bochum, Chair of Work, Human Resource & Leadership and Coordinator of the Competence Center HUMAINE and the HUMAINE Network e.V., an Association for Human-Centered AI.

Address: Ruhr University Bochum, Institute of Work Science, Chair of Work, Human Resource & Leadership, Universitätsstr. 150, 44801 Bochum, Germany,
E-Mail: niklas.obermann@rub.de
ORCID: https://orcid.org/0009-0004-3817-3203

Daniel Lupp, Dr. rer. oec., contributed to this article until summer 2025 while he was a Research Associate at Ruhr University Bochum. He received his Ph.D. from the Faculty of Management & Economics at Ruhr University Bochum and now works as a Senior Consultant at an international consulting firm. He contributed equally to the third author (alphabetical order).

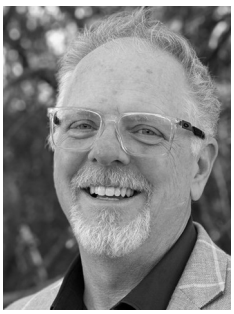
Address: Ruhr University Bochum, Institute of Work Science, Chair of Work, Human Resource & Leadership, Universitätsstr. 150, 44801 Bochum, Germany,
E-Mail: daniel.lupp@rub.de
ORCID: https://orcid.org/0000-0001-5925-0598

Uta Wilkens, Prof. Dr., is Full Professor of Work, Human Resources & Leadership at Ruhr University Bochum, Germany, Member of acatech "*Learning Systems*", Chairperson of the Competence Center HUMAINE and the HUMAINE Network e.V., an Association for Human-Centered AI. She contributed equally to the second author (alphabetical order).

Address: Ruhr University Bochum, Institute of Work Science, Chair of Work, Human Resource & Leadership, Universitätsstr. 150, 44801 Bochum, Germany,
E-Mail: uta.wilkens@rub.de
ORCID: https://orcid.org/0000-0002-7485-4186

Received: 29.09.2025 | Revised: 11.01.2026 | Accepted: 15.01.2026

Trust and Responsibility in AI: An Interdisciplinary Social-Sector Perspective



Nathan Chappell

Abstract: The rapid adoption of artificial intelligence has intensified debates about responsibility, ethics, and trust. While regulatory frameworks and organizational ethics statements are proliferating, responsible AI is too often treated as compliance or reputation management rather than an organizing principle of practice. This perspective argues that social-sector organizations - including nonprofits, NGOs, and other mission-driven institutions - offer an instructive lens for rethinking responsible AI because they operate with structural vulnerability and high trust dependence. Drawing on business ethics, organizational theory, nonprofit and social-sector management, and human flourishing scholarship, the paper proposes shifting from harm-avoidance toward trust-centered, flourishing-oriented AI integration.

Keywords: Responsible AI; Trust; Social sector; Nonprofit management; Human flourishing

Vertrauen und Verantwortung in der KI: Eine interdisziplinäre Perspektive aus dem sozialen Sektor

Zusammenfassung: Die rasante Verbreitung künstlicher Intelligenz hat die Debatten über Verantwortung, Ethik und Vertrauen intensiviert. Während regulatorische Rahmenbedingungen und ethische Leitbilder von Organisationen immer zahlreicher werden, wird verantwortungsvolle KI allzu oft als Compliance- oder Reputationsmanagement behandelt und nicht als organisierendes Prinzip der Praxis. Diese Perspektive argumentiert, dass Organisationen des sozialen Sektors - darunter gemeinnützige Organisationen, NGOs und andere missionsorientierte Institutionen - eine lehrreiche Perspektive für ein Umdenken in Bezug auf verantwortungsvolle KI bieten, da sie mit struktureller Vulnerabilität und hoher Vertrauensabhängigkeit arbeiten. Auf der Grundlage von Wirtschaftsethik, Organisationstheorie, Management im gemeinnützigen und sozialen Sektor sowie wissenschaftlichen Erkenntnissen zur menschlichen Entfaltung schlägt der Artikel vor, von der Vermeidung von Schaden hin zu einer vertrauenszentrierten, auf menschliches Wohlergehen ausgerichteten KI-Integration überzugehen.

Stichwörter: Verantwortungsvolle KI; Vertrauen; Sozialer Sektor; Nonprofit-Management; Menschliche Entfaltung

Introduction - Responsibility or Reputation?

The rapid adoption of artificial intelligence across industries has sparked urgent debates about responsibility. In the corporate world, however, responsible AI too often risks collapsing into reputation management. Declarations of ethical intent, ESG statements,

and glossy cause marketing campaigns abound, yet responsibility is frequently treated as a safeguard for brand image rather than an organizing principle of practice. The consequences are not hypothetical. Biased algorithms and opaque systems already show how easily efficiency eclipses fairness when responsibility is an afterthought.

The social sector offers a strikingly different lens. For nonprofits, responsibility is not optional, nor is it a reputational exercise. It is existential. Their missions depend on fragile bonds of trust with donors, beneficiaries, and communities. Without that trust, funding dries up, partnerships dissolve, and impact is curtailed. This makes nonprofits a vital proving ground for what responsible AI should look like.

As both a nonprofit practitioner and an architect of AI solutions, my vantage point is shaped by years of seeing how technology collides with mission, values, and human dignity. From this perspective, the nonprofit sector is both a warning system and a blueprint for what responsible AI must become.

Structural Vulnerability as Clarity: A Social-Sector Lens

Many organizations adopt AI with powerful incentives to scale quickly, improve efficiency, and capture competitive advantage. In large enterprises, formal strategies and secured software landscapes can accelerate adoption, sometimes allowing responsibility to drift into a form of "risk management" handled after value has been extracted. At the same time, in small and medium-sized enterprises (SMEs), family businesses, start-ups, and social-sector organizations, adoption is often far less centralized - employees bring their own devices, experiment with generative tools, and use software without an official strategy or license. This "shadow AI" reality can increase both productivity and exposure, especially when data governance, privacy expectations, and accountability are unclear.

Social-sector organizations, by contrast, often operate with structural vulnerability. Resource scarcity is not unique to nonprofits - many private companies, especially SMEs and start-ups, also face tight margins. What differs is how quickly trust loss becomes existential for mission-driven institutions. A single ethical lapse - such as a biased allocation of resources or a misstep in data privacy involving vulnerable communities - can permanently damage credibility with donors, beneficiaries, partners, and regulators. In many social-sector contexts, there is no practical second chance to "pivot" after a failed experiment. The risks are immediate and relational.

Yet scarcity is not only a constraint; it functions as a kind of ethical clarity. Decisions about whether and how to adopt AI are filtered through a sharper lens: Will this tool uphold or erode trust? Will it enhance dignity or compromise it? With limited resources, nonprofits must scrutinize technology not for how it scales operations, but for how it aligns with mission. The absence of financial incentives enables nonprofits to see risks and opportunities more clearly.

By being forced to ask first-order ethical questions rather than defaulting to market-driven metrics, nonprofits reveal both the dangers of unchecked AI adoption and the possibility of aligning innovation with responsibility. Scarcity produces not paralysis but vision - the ability to see the forest through the trees.

Trust as the True Currency

If corporations often measure success in profits, social-sector organizations ultimately measure success in impact - with trust as the enabling condition that makes sustained impact possible. Every donation and partnership rests on confidence that resources are stewarded responsibly. In the context of AI, this means that ethical lapses are not minor reputational setbacks; they are existential threats that can unravel years of relationship-building in an instant.

Nonprofit adoption of AI - whether in donor engagement, program delivery, or operations - is therefore filtered through the trust imperative. Tools must not only function effectively but also be explainable, transparent, and aligned with mission values. A predictive system that improves efficiency while undermining fairness or privacy is unacceptable because it erodes nonprofit legitimacy.

Nonprofits rarely have that luxury. Once trust is broken, donors move on, beneficiaries lose confidence, and the organization may never recover. This asymmetry makes nonprofits uniquely attuned to the ethical stakes of AI.

From this vantage point, responsible AI is not merely about avoiding harm. It is about sustaining the fragile bonds of confidence between institutions and the people they serve. For corporations seeking to embed ethics into practice, the nonprofit example offers a living model of trust-centered AI integration - an approach where responsibility is not a public-relations veneer but an existential requirement.

From Standards to Culture

The European Union's AI Act (European Union, 2024), with its tiered risk classifications, represents the most comprehensive regulatory effort to date. UNESCO's Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021) and the OECD's AI Principles (Organisation for Economic Co-operation and Development [OECD], 2019) further articulate global commitments to fairness, transparency, and accountability. Alongside these, ESG frameworks and corporate declarations of AI ethics proliferate. Together, these instruments provide an important set of guardrails, creating a shared vocabulary for what responsible AI should mean in practice.

Yet a persistent tension remains: standards are necessary but insufficient. A company can comply fully with existing regulations while still treating responsibility as a box to check rather than a value to embody. The danger is compliance theater - satisfying reporting requirements without cultivating a culture of responsibility. In such cases, ethics is outsourced, and responsibility reduced to paperwork.

Their limited resources rarely allow for elaborate compliance structures or dedicated ethics offices. Instead, responsibility must be integrated into the day-to-day culture of the organization because survival demands it. Decisions about whether to use AI for something as simple as scheduling volunteers are not purely operational; they are ethical. Is it accessible, fair, and transparent? These questions cannot be outsourced; they must be lived daily.

For businesses navigating the complexities of responsible AI, this social sector perspective is instructive. The challenge is not merely harmonizing diverse standards or reporting systems. It is embedding responsibility into the cultural DNA of organizations, where ethical reflection informs not just what companies report, but how they operate. Responsibility, in this sense, must move from external compliance to internal conviction.

Beyond Ethics: Human Flourishing as the Horizon

Ethics should be the minimum expectation for AI. The more urgent and transformative question is whether AI contributes to human flourishing. Nonprofits are uniquely positioned to surface this broader horizon because their missions are already aligned with advancing human well-being.

Human flourishing can be understood across seven empirically grounded dimensions - character, health, relationships, finances, happiness, faith, and meaning (VanderWeele, 2017). Together, they reflect a view of well-being that extends beyond compliance. Each dimension offers a lens for evaluating AI not only for its risks but also for its capacity to enhance life. AI that reinforces fairness strengthens character; systems that protect access to care or reduce unnecessary barriers advance health; technologies that foster inclusion enrich relationships. When AI reduces inequality, it supports financial dignity; when it lowers stress or creates space for joy, it contributes to happiness.

These dimensions sharpen the contrast between "do no harm" and "do good." Corporate responsibility often frames ethics in terms of preventing harm - avoiding bias, ensuring compliance, protecting privacy. While necessary, this bar is too low. Flourishing asks a deeper question: does this technology make life more worth living? For nonprofits, the answer resonates naturally because their missions already align with these outcomes. For corporations, adopting a flourishing-centered framework represents a paradigm shift, moving responsible AI out of the realm of marketing or regulation and into the realm of purpose.

This shift changes governance conversations from "Are we compliant?" to "Are we worthy of trust, and does this system make life more worth living for the people touched by it?"

Implications for Future Research

Responsible AI in trust-dependent and mission-driven contexts raises several research questions that merit further inquiry:

1. How can organizations operationalize trust as a governance and performance measure without reducing it to an instrumental variable?
2. Which governance models best support flourishing-oriented AI adoption across sectors, including SMEs and social-sector organizations?
3. How do AI-related ethical failures differ structurally and relationally across organizational types, and what recovery pathways exist?
4. What practices most effectively embed responsibility into organizational culture rather than delegating it to compliance functions?
5. How can human flourishing be operationalized as an evaluative standard for AI systems, and what indicators or metrics could credibly assess an AI model's contribution to individual and collective well-being?

Conclusion – Learning from the Canaries

Nonprofits are the canaries in the coal mine of AI ethics. Their scarcity forces clarity, their survival depends on trust, and their missions are aligned with flourishing. Unable to treat ethics as marketing or an add-on, nonprofits reveal both the perils of neglect and the

promise of alignment. They show how easily dignity is compromised when technology is careless, and how powerfully well-being grows when responsibility is embedded.

For corporate actors, the lesson is unmistakable. Responsible AI cannot remain an external exercise in reputation or compliance. It must become a cultural commitment - a way of doing business inseparable from identity. Standards and regulations may prevent the worst abuses, but they cannot create authentic responsibility. That requires conviction, leadership, and a willingness to adopt new measures of success.

If ethics is the floor, flourishing is the ceiling. The nonprofit sector demonstrates how this shift is possible: by making trust the central metric, embedding responsibility into culture, and aligning technology with humanity's deepest aspirations. Businesses that learn from these canaries will not only avoid catastrophe; they will help chart a path toward an AI future where innovation and flourishing reinforce one another.

The coal mine of AI adoption is fraught with risk, but the canaries are already singing.

References

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People – An ethical framework for a good AI society: Opportunities and risks. Minds and Machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709-734. https://doi.org/10.5465/amr.1995.9508080335

Organisation for Economic Co-operation and Development. (2019). OECD AI principles. Adopted May 2019; updated May 2024. OECD. https://oecd.ai/en/ai-principles

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. Adopted 23 November 2021. United Nations Educational, Scientific and Cultural Organization.

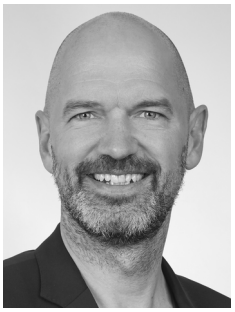
VanderWeele, T. J. (2017). On the promotion of human flourishing. Proceedings of the National Academy of Sciences of the United States of America, 114(31), 8148-8156. https://doi.org/10.1073/pnas.1702996114

Nathan Chappell is Chief AI Officer at Virtuous and a leading voice at the intersection of generosity, responsible AI, and social impact. He is the author of Nonprofit AI: A Comprehensive Guide to Implementing Artificial Intelligence for Social Good and The Generosity Crisis: The Case for Radical Connection to Solve Humanity's Greatest Challenges. His work focuses on trust-centered AI adoption, ethical innovation, and the application of advanced technologies to strengthen human connection and social good.

Address: Virtuous Software, 1 N 1st St, Phoenix, AZ 85004, USA.
E-mail: nathan.chappell@virtuous.org
ORCID: https://orcid.org/0009-0006-2487-6332

Received: 17.12.2025 | Revised: 22.01.2026 | Accepted: 29.01.2026

Steps Towards "Responsible" and Human-Centered "AI" – Some Ethical Considerations



Peter G. Kirchschlaeger

Abstract: So-called "artificial intelligence (AI)" – more adequately referred to as "data-based systems (DS)" – opens ethical downsides and upsides. There is a necessity to identify precisely the ethical opportunities and risks of DS in order to promote the former and in order to avoid the latter – for the benefit of all people and the planet earth. Companies can contribute to the realization of DS with ethics by, first, living up to the exclusive human responsibility for machines; second, while running innovation- and research-processes, by implementing always right from the start an interaction between ethics and technologies; third, by promoting global human rights-based regulation of DS as well as the establishment of an International Data-Based Systems Agency (IDA) at the UN enforcing this global regulation of DS.

Keywords: Artificial Intelligence (AI), Data-Based Systems (DS), Ethics, Human Rights, International Data-Based Systems Agency (IDA), UN

Schritte zu einer "verantwortungsvollen" und menschenzentrierten "KI" – ein paar ethische Überlegungen

Zusammenfassung: Die sogenannte „künstliche Intelligenz (KI)" – besser bezeichnet als „datenbasierte Systeme (DS)" – birgt ethische Vor- und Nachteile. Es ist notwendig, die ethischen Chancen und Risiken von DS genau zu identifizieren, damit die ethischen Vorteile gefördert und die ethischen Nachteile der DS vermieden werden können – zugunsten aller Menschen und des Planeten Erde. Unternehmen können zur Verwirklichung ethischer DS beitragen, indem sie erstens der ausschliesslichen Verantwortung des Menschen für Maschinen gerecht werden, zweitens bei der Durchführung von Innovations- und Forschungsprozessen von Anfang an eine Interaktion zwischen Ethik und Technologien implementieren und drittens eine globale, auf Menschenrechten basierende Regulierung von DS sowie die Einrichtung einer Internationalen Agentur für datenbasierte Systeme (IDA) bei der UNO fördern, die diese globale Regulierung von DS durchsetzt.

Schlüsselwörter: Künstliche Intelligenz (KI), Datenbasierte Systeme (DS), Ethik, Menschenrechte, Internationale Agentur für datenbasierte Systeme (IDA), UNO

1 So-Called "AI"? Data-Based Systems (DS)!

The ethical analysis of so-called "Artificial Intelligence (AI)" starts with a critical examination of the term itself from an ethical standpoint. So-called "AI" can be defined as striving by technical means to imitate or fulfill cognitive functions of human thought. During an ethical critique of so-called "AI", it becomes clear that so-called "AI" does not comprise

the sum of human knowledge, nor is it objective, fair and neutral. It is not trained by or does not represent reality, facts, or scientific evidence, but is trained by and represents the online past. It is only based on certain data. Looking at the above-mentioned definition of so-called "AI" from an ethical perspective, the term "artificial" is not questioned because this technology is created by humans.

A first criticism arises regarding "intelligence" because intelligence does not just consist of the solution of a cognitive task but also in the way it is pursued (Misselhorn, 2018). A second criticism highlights that so-called "AI" is limited to certain areas of intelligence (e.g., certain cognitive capacities). Among others, in the domain of emotional and social intelligence, machines are only able to simulate emotions, personal interaction, and relationships and lack authenticity. For instance, a health care robot can be trained to cry when the patient is crying, but no one would argue that the robot feels real emotions and cries due to them. The robot does not even care nor not care about it. On the contrary, one could train the exact same robot to slap the patient's face when the patient is crying, and the robot would perform this function in the same perfect way. Again, the robot does not even care nor not care about it. Machines cannot reach emotional and social intelligence – neither today nor tomorrow as the expectable further increase of compute of machines in the future can indeed improve the simulation of emotions by machines but does not create emotional and social intelligence.

Beyond that area of human intelligence, in the domain of moral capability, one cannot ascribe machines with moral capability because they are presupposed to follow ethical rules given by humans. Technologies are primarily made for their suitability and may set rules as a self-learning system, for example, to increase their efficiency, but these rules do not contain any ethical quality. E.g., a self-driving car could set the rules for itself, but it is not aware of the ethical quality of these rules. It could give itself the rule to get from A to B as fast as possible including harming humans and nature, to optimally fulfill the task of reaching B in the shortest time possible, without being able to recognize ethical rules for itself, which would allow the machine to perceive the illegitimacy of its rules and actions. A human driver instead possesses the potential to recognize for himself or herself binding ethical rules, which empower him or her to see that harming humans and nature might be more efficient but illegitimate. While humans are able to recognize by themselves ethical rules for themselves (Kirchschlaeger, 2023), machines cannot. *They don't even do not care* if they fulfill a legitimate or illegitimate task. The potential that DS possess in relation to ethical actions is nowhere close to moral capability because DS lack not only autonomy but also vulnerability, conscience, freedom, and responsibility, which are all essential for human morality (Kirchschlaeger, 2021).

The term "data-based systems (DS)" (Kirchschlaeger, 2021; 2022) would be more appropriate than "AI" because this term describes what actually constitutes "AI": generation, collection, and evaluation of data; data-based perception (sensory, linguistic); data-based predictions; data-based decisions. The mastery of an enormous quantity of data depicts the key asset of these technologies – of *data-based systems (DS)*.

The above reflection leads to the main conclusion that DS can not be responsible. Humans are and remain exclusively responsible for DS (Johnson, 2006; Yampolski, 2013). Companies need to live up to this exclusive responsibility for DS ensuring that DS are human-centered.

2 Implementing Interaction Between Ethics and Technologies

Beyond that, companies should avoid understanding ethics as an afterthought in ventures and innovation- and research-processes (Kirchschlaeger, 2024a; 2024b). Instead, the relationship of ethics and technology should be understood as an interaction, with each contributing to the other (Kirchschlaeger, 2021). Applications of groundbreaking technologies often reshape the ethical environment by creating new solutions to societal challenges and new values. At the same time, scientists and technologists all perform their work within an ethically informed context. Meanwhile, ethics contributes to technology by stimulating technological innovation, by recognizing technological inventions, and by providing ethical guidance.

Moreover, ethics belongs to technology. Horizons of meaning and ethical ends inform technology in an ethical sense. Ethics should be considered right from the start because of the very nature of technology as a human creation. Ethics can provide ethical guidance to the agenda-setting for innovation and research.

Finally, ethical principles and norms inform legal principles and norms guaranteeing freedom and independence of research. Only this freedom and this independence enable new explorations and insights as well as innovation.

3 Promoting Human Rights-Based Global Regulation and the Establishment of an International Data-Based Systems Agency (IDA) at the UN

Humans need to become active so that DS do not simply happen, but that humans shape them. This is necessary so that DS will not be reduced to an instrument serving pure efficiency but can rise to their ethically positive potential. More importantly, there is a need for ethical guidance to review the economic self-interests that run DS so far almost exclusively.

Beyond the so-far elaborated two concrete measures on an organizational level, living up to the exclusive human responsibility for DS as well as the promotion of an interaction between ethics and technologies right from the start of a venture or a research- and innovation-process (meso-level), companies should – on a macro-level – join the efforts striving for human rights-based DS (a global regulatory framework encompassing the respect and implementation of human rights in the entire life-circle of DS) (Kirchschlaeger, 2021; 2024c; 2025) and support the establishment of an International Data-Based Systems Agency (IDA) at the UN (Kirchschlaeger, 2021). IDA at the UN should fulfill the following three key functions:

1. Providing an access to market approval-process which several other industries know since decades (e.g., the pharmaceutical industry) in order to avoid harm of humans and the environment; the access to market approval-process orchestrated by IDA should ensure that human rights-violating DS including so-called "frontier models" as well as applications and products (like, e.g., an app sexualizing pictures of children) (Heikkilae, 2022; Snow, 2022; Lensa, 2025) do not even end up on the market. By this, a preventive impact is caused by IDA that the private sector does not even design and develop such human rights-violating DS knowing that they will not pass the access to market approval-process;
2. Monitoring that human rights are not violated with or by DS;

3. Fostering international technical collaboration in the sphere of DS in order to enable humanity to reach faster and better the positive potential of DS.

The aim of IDA at the UN is to ensure and to promote the development and deployment of HRBDS as a regulatory framework guaranteeing the use of the ethical positive potential of DS for the benefit of all humans and the planet as well as the handling of its ethical negative potential, including the destruction of humankind and the planet. Serving this aim is the establishment of robust governance mechanisms.

IDA should be built following the model of the International Atomic Energy Agency (IAEA) at the UN as an "institution with teeth" because, thanks to its legal powers, functions, enforcement mechanisms, and instruments, the IAEA was able to foster innovation and ethical opportunities while at the same time protecting humanity and the planet from the existential risks in the domain of nuclear technologies, which also embrace the same dual nature as DS, covering both ethical upsides and downsides. Leveraging the lessons learned from nuclear technologies and the establishment of the IAEA, the establishment of IDA presents a viable pathway towards effective global governance of existential AI risks, ensuring the responsible and ethical development of DS for the betterment of humanity and the planet.

What makes the establishment of an IDA realistic is not only its essential and minimum normative framework, its practice-oriented and participatory governance-structure, , as well as its striving for legitimacy combined with fostering innovation but also that in the past, humanity has shown that when the well-being of people and the planet is at stake, humanity can focus on what is technically feasible rather than blindly pursuing all that is technically possible.

Humanity did pursue nuclear technology, develop the atomic bomb, and even deploy it more than once. But to prevent yet worse events, humanity then massively restricted the research and development of nuclear technology despite overwhelming opposition by state and non-state actors. That nothing worse has happened is largely due to international guidelines, concrete enforcement mechanisms, and the International Atomic Energy Agency (IAEA) of the UN (Kirchschlaeger, 2021).

Beyond that, DS distinguish themselves from nuclear technology especially in three characteristics that increase the realizability of the establishment and the existential impact of IDA for humanity and the planet:

1. In order to function, DS must have power. This means that if a DS is violating human rights, threatening peace, or destroying the planet, it can be stopped by taking it off the power grid or by cutting off the power supply.
2. In order to function, DS must be connected because of its dependence on data flow. This means that if a DS is violating human rights, threatening peace, or destroying the planet, it can be stopped by disconnecting it.
3. While operating, every DS leaves data traces, allowing identification and accountability.

Finally, it also builds an advantage over attempts to ensure that all humans benefit from a previous technology-based innovation and to master the ethical dangers of a previous technology-based innovation (like in the case of nuclear technologies with the establishment of the International Atomic Energy Agency), IDA could also rely on DS-based solutions to implement HRBDS.

Supporting HRBDS and IDA means for companies to join forces with a fast growing international and interdisciplinary network of experts, business leaders, entrepreneurs, and global leaders (IDA, 2025), including, among others, UN Secretary General António Guterres, Pope Francis, His Holiness the Dalai Lama, UN High Commissioner for Human Rights Volker Türk, Sam Altman (Founder of Open AI), Mustafa Suleyman (CEO of Microsoft AI, Co-Founder and former Head of applied AI at DeepMind), and "The Elders" (Kirchschlaeger, 2024b; 2024d; 2024e; 2024f; 2024g; 2025).

References

Heikkilae, M. (2022, December 12). The viral AI avatar app Lensa undressed me – without my consent. MIT Technology Review. https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent

IDA. (2025). International Data-Based Systems Agency IDA at the UN: Supporters of IDA. IDA. Retrieved October 9, 2025, from https://idaonline.ch/supporters-of-ida/

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. Ethics and Information Technology, 8(4), 195-204. https://doi.org/10.1007/s10676-006-9111-5

Kirchschlaeger, P. G. (2021). Digital Transformation and Ethics. Ethical Considerations on the robotization and automation of society and the economy and the use of Artificial Intelligence. Nomos.

Kirchschlaeger, P. G. (2022). Ethische KI? Datenbasierte Systeme (DS) mit Ethik. HMD-Praxis der Wirtschaftsinformatik, 59(2), 482-494. https://doi.org/10.1365/s40702-022-00843-2

Kirchschlaeger, P. G. (2023). Ethical Decision-Making. Nomos. https://doi.org/10.5771/9783748918684

Kirchschlaeger, P. G. (2024a, April 11). In an era of digital disruptions, ethics can't be an afterthought – Part 1. Business and Human Rights Journal Blog.

https://www.cambridge.org/core/blog/2024/04/11/in-an-era-of-digital-disruptions-ethics-cant-be-an-afterthought/

Kirchschlaeger, P. G. (2024b, April 12). In an era of digital disruptions, ethics can't be an afterthought – Part 2. Business and Human Rights Journal Blog. https://www.cambridge.org/core/blog/2024/04/12/in-an-era-of-digital-disruptions-ethics-cant-be-an-afterthought-part-2/

Kirchschlaeger, P. G. (2024c). Artificial intelligence and the complexity of ethics. Asian Horizons, 14(3), 375-389. https://dvkjournals.in/index.php/ah/article/view/4590/3752

Kirchschlaeger, P. G. (2024d, December 3). Protecting children from Anti-Social media. *Project Syndicate*. https://www.project-syndicate.org/commentary/australia-ban-on-children-using-social-media-should-be-emulated-by-peter-g-kirchschlager-2024-12

Kirchschlaeger, P. G. (2024e). The need for an International Data-Based Systems Agency (IDA) at the UN: governing "AI" globally by keeping the planet sustainably and protecting the weaker from the powerful. Journal of AI Humanities, 18, 213-248.

Kirchschlaeger, P. G. (2024f). An International Data-Based Systems Agency IDA: striving for a peaceful, sustainable, and Human Rights-Based future. Philosophies, 9(3), 73. https://doi.org/10.3390/philosophies9030073

Kirchschlaeger, P. G. (2024g). Securing a peaceful, sustainable, and humane future through an International Data-based Systems Agency (IDA) at the UN. Data & Policy, 6(78). https://doi.org/10.1017/dap.2024.38

Kirchschlaeger, P.G. (2025). Artificial Intelligence – an Analysis from the Rights of the Child Perspective. Berkley Journal of International Law. https://www.berkeleyjournalofinternationallaw.com/post/artificial-intelligence-an-analysis-from-the-rights-of-the-child-perspective

Lensa. (2025). Lensa AI: Influencers' best kept secret. Lensa App. Retrieved October 9, 2025, from https://lensa.app/

Misselhorn, C. (2018). Grundfragen der Maschinenethik. Reclam.

Snow, O. (2022, December 7). 'Magic Avatar' app Lensa generated nudes from my childhood photos. The dreamy picture-editing AI is a nightmare waiting to happen. Wired. https://www.wired.com/story/lensa-artificial-intelligence-csem/?bxid=5cc9e15efc942d13eb203f10

Yampolskiy, R.V. (2013). Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach. In V. Müller (Ed), Philosophy and Theory of Artificial Intelligence. Studies in Applied Philosophy, Epistemology and Rational Ethics (pp. 389-396). Springer. https://doi.org/10.1007/978-3-642-31674-6_2

Peter G. Kirchschlaeger, Prof. Dr., is Ethics-Professor and Director of the Institute for Social Ethics ISE at University of Lucerne, Research Fellow at the University of the Free State, Bloemfontein (South Africa), Visiting Professor at the Chair of Neuronal Learning and Intelligent Systems at ETH Zurich and at the ETH AI Center as well as Visiting Fellow at the University of Tuebingen (Germany). Previously, he was a Visiting Fellow at Yale University (USA).

Address: University of Lucerne, Institute of Social Ethics ISE, Frohburgstrasse 3, Postfach, 6002 Luzern, Switzerland, Tel.: +41 41 229 52 61, E-Mail: peter.kirchschlaeger@unilu.chORCID: https://orcid.org/0000-0001-9528-1228

Received: 21.09.2025 | Revised: 18.01.2026 | Accepted: 29.01.2026

Defining Knowledge in the Age of Society 5.0

Susanne Durst

Abstract: The aim of this perspective contribution is to introduce the concept of responsible knowledge management (rKM) and its usefulness for implementing ethically accepted AI solutions in organizations. Illustrative examples are presented to demonstrate the latter. The article concludes with a series of research questions that serve as an outlook and inspiration for further reflection on rKM and ethically acceptable AI solutions in companies.

Keywords: Responsible AI, human-centred AI, responsible knowledge management, ethical behaviour, responsible entrepreneurship, inclusivity

Definition von Wissen im Zeitalter der Gesellschaft 5.0

Zusammenfassung: Ziel dieses Perspektivbeitrags ist es, das Konzept des verantwortungsvollen Wissensmanagements und dessen Nutzen für die Implementierung ethisch akzeptierter KI-Lösungen in Organisationen vorzustellen. Letzteres wird anhand anschaulicher Beispiele verdeutlicht. Der Artikel schließt mit einer Reihe von Forschungsfragen, die als Ausblick und Inspiration für weitere Überlegungen zum verantwortungsvollen Wissensmanagement und ethisch vertretbaren KI-Lösungen in Unternehmen dienen sollen.

Stichwörter: Verantwortungsvolle KI, menschenzentrierte KI, verantwortungsvolles Wissensmanagement, ethisches Verhalten, verantwortungsvolles Unternehmertum, Inklusion

Responsible and human-centred AI in organizations and knowledge

The development and implementation of regulations, declarations and standards to ensure the responsible design and introduction of artificial intelligence (AI), as well as the ethically acceptable integration of AI into existing business processes, pose considerable challenges for companies (Mökander et al., 2022; Schneider et al., 2023). This applies to all sort of companies, young and mature companies, small and large enterprises. The associated activities and behaviours not only require the mobilisation of existing knowledge but also suggest the development of new knowledge while consciously forgetting or unlearning old knowledge (de Holan & Phillips, 2004) in order to be ready for informed decisions and subsequent next steps. This knowledge development or updating of existing knowledge should take place across companies, and consequently every single member of the company must be included and convinced.

This, in turn, underscores once more the importance of having a dedicated and systematic approach to knowledge management (KM) in companies (Zack, 2002). Davenport's classic definition describes KM as "the process of capturing, distributing, and effectively using knowledge" (Davenport, 1994). However, from the author's point of view, the underlying notions of "traditional" KM thinking has its limitations when it comes to

responsible and human-centred AI in organizations because it is still primarily aimed at private companies and the creation and maintenance of competitive advantages (Davenport & Prusak, 1998; Alavi & Leidner, 2001; Ferreira et al., 2020), and above all, it is primarily seen as something positive, as an asset (Caddy, 2000; Durst & Zieba, 2019). This way of thinking appears to be too narrow to anchor the goal of ethically acceptable integration of AI into companies' business processes and models. If knowledge is viewed solely as something valuable, there is a risk of overlooking the many situations in which this knowledge is more like a liability. Rigorous and sound debates about responsible AI in corporate business ethics, however, require a more nuanced use and perception of knowledge, especially its origin, to strengthen organizations' confidence in the identified and selected solutions.

Therefore, the author of this perspective paper argues that the development and consequences of AI in and for organizations make it necessary to view knowledge from a broader perspective; a perspective that goes beyond that of the individual organization and refrains from viewing knowledge per se as something positive (Durst & Foli, 2024). As a possible solution, the author proposes the concept of responsible knowledge management (rKM).

Responsible knowledge management as a way forward achieving responsible and human-centred AI in organizations?

RKM focuses on responsible KM processes such as the creation, transfer, preservation and application of knowledge for the common good, i.e., for measures that benefit society as a whole or contribute to a better society. RKM has people in the core, i.e., technology is (and remains) subordinate to people. This approach to KM is human-centred, inclusive and collaborative and invites everyone to contribute but also to take responsibility. Consequently, rKM involves different and diverse partners on an equal footing (Durst, 2024). Inclusivity takes an active role in rKM in all activities and throughout the rKM lifecycle (Dalkir, 2025). Additionally, rKM incorporates Dyllick and Muff's (2016) typology of business sustainability, meaning that the starting point of organizations and their actions and behaviour is the external environment. This means that a society-centred approach is emphasized (Durst & Khadir, 2025), which also connects rKM with the concept of Society 5.0, which places people at the centre of innovation (Carayannis & Morawska-Jancelewicz, 2022). The importance of knowledge for innovation is well known (Du Plessis, 2007). This can also be applied to innovative solutions for the development and implementation of ethically acceptable AI in businesses.

Knowledge, from a rKM perspective, is seen as something neutral and, depending on the situation, it can be positive, negative or both. If one adopts the language of the insurance industry, this means that knowledge can be seen as a positive and/or negative risk, depending on the situation. To illustrate this from a knowledge at risk perspective (Williams & Durst, 2019), imagine the departure of a long-standing employee. Negative: The employee's departure leads to an attrition of knowledge or even a loss of knowledge. Neutral: The departure has no consequences for the company or the direct colleagues of the departing employee. Positive: The flow of knowledge is improved, as more knowledge is now shared and integrated into the company's knowledge processes.

The following examples attempt to demonstrate the advantages of a neutral approach to knowledge when the goal is the ethically acceptable integration of AI into existing busi-

ness processes. For example, there could be a knowledge gap between what the company should know (or should be able to know) to implement AI ethically and what it actually knows. A knowledge gap also represents a clear signal that entirely new knowledge must be developed and/or old knowledge updated. This process requires a critical discussion of existing knowledge, which means that it must first be identified and then reviewed for its relevance to addressing the current challenge, i.e., for developing an ethically acceptable AI solution within the company. The result of this review could be that the existing knowledge is insufficient, and new knowledge must be created. However, achieving this may require a conscious process of unlearning (McGill & Slocum, 1993) as a basis for creating a new framework/mindset necessary for the implementation of AI solutions and their expected impact on the organization.

Another example could be that decisions are made quickly and without careful consideration, as a result of reacting to external pressure—for example, competitors rush ahead, or ecosystem partners, especially dominant players, try to force certain AI solutions on the company. Certain media reports and their chosen connotations when discussing the importance of AI implementations in companies to remain competitive could also lead to hasty decisions. Especially if these decisions are associated with an unreflective belief in technology, in the sense that AI is the panacea for fundamental problems in the company and/or the world. In other words, the company is not using its own knowledge and judgment to make an informed decision. Conversely, a state of denial can be considered equally critical. However, it should also be noted that decisions are also made on the basis of underdeveloped, missing, outdated, unreliable or incorrectly applied knowledge (Durst & Zieba, 2018).

It is conceivable that the implementation of AI solutions in a company will require collaboration with new service providers who provide and maintain the technical solutions (Merhi, 2023). At the same time, it is quite likely that the ethical principles of the business partners differ. However, in order to develop and subsequently implement ethically acceptable AI solutions within a company, the partners must first clarify whether their values, norms, and ideas are compatible before examining the specific solution in more detail. If this is not the case, it is advisable to seek an alternative partner, as the company's ethical values and norms should be non-negotiable. The same approach would also apply to existing information and communications technology (ICT) partners.

The core idea is that those who are prepared to view their own knowledge and the collective knowledge of the company from a neutral perspective should be better able to consider both the positive and negative consequences of a project and discuss them with others. This, in turn, should also benefit their judgement. Consequently, rKM recognises the value and importance of risk management and risk management literacy among people. Knowledge in Society 5.0, and thus rKM, must also be supported by AI literacy, which is fundamental not only for the ethical and responsible use of AI, but also for its development and implementation (Ng et al., 2021).

The discussion of ethically acceptable AI solutions, their selection, and subsequent implementation requires not only a diverse range of relevant and reliable knowledge (old and new), but also that this knowledge is contributed and (constantly) critically examined by different knowledge holders. A company that not only welcomes diversity and inclusion but actually practices it brings everyone to the table, regardless of gender, cultural or professional background, or position within the company. Such a company gives everyone

a voice, but at the same time emphasizes that this voice also comes with responsibility. This also means that leadership is developed (negotiated) in the discussions and thus represents a result and is not something that is pre-existing due to a position of authority.

Let's try "responsible knowledge management" thinking

The aim of this perspective paper was to introduce the concept of rKM to offer an interesting framework of KM for tackling the implications of ethically acceptable AI integrations in organizations. Integrating the underlying principles of rKM into discussions related to responsible and human-centred AI in business ethics is expected to lead to the development and execution of more inclusive and responsible solutions to addressing the challenges at hand. This way of thinking can also contribute to achieving the United Nations Sustainable Development Goals, in particular Goal 5 "Gender Equality," Goal 10 "Reduced Inequalities," and finally Goal 17 "Partnerships." The use of AI raises the question of digital inequality, which is why a human-centred, inclusive, and collaborative approach such as rKM is more important than ever.

The author is aware that the rKM concept is challenging and that some people will question its actual feasibility. However, if ethical behaviour and actions for and within organizations are truly considered important, and are not just lip service, one can (and must) expect to be confronted with some "unpleasant" challenges as an individual. To move from thinking to action, the following questions can serve as food for thought:

What are the factors that enable responsible KM thinking that responds to the current activities and initiatives of organizations to act and work ethically and responsibly?

What forms of organization and/or ownership structures would be needed to promote/enable rKM thinking?

How to make sure that organizations take sufficient time to think things through thoroughly before making hasty decisions, despite the rapid development of new AI models and solutions?

Since involving different people and their views and ideas requires tolerance of ambiguity, the question arises as to how one can develop such a skill, considering that dealing with ambiguity is frightening for many people. How should the school and education system change in this regard?

With all the planned initiatives regarding the introduction of ethically acceptable AI measures, organizations must also be prepared for their failure. What can be done to ensure that efforts are not abandoned in the face of the stakes?

Responsible knowledge management requires a willingness to take responsibility on the part of all those involved. How can the field of leadership research, particularly the subfield of self-leadership, contribute to the realisation of this principle?

The concept of good and bad actions and behaviour varies from person to person. How can an organization ensure that its members agree on a solution? And once this compromise has been reached, how can it be ensured that it is incorporated into the selection and introduction of ethically acceptable AI solutions into existing business processes?

References

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. MIS Quarterly, 25(1), 107–136. https://doi.org/10.2307/3250961

Caddy, I. (2000). Intellectual capital: recognizing both assets and liabilities. Journal of Intellectual Capital, 1(2), 129-146.

Carayannis, E. G., & Morawska-Jancelewicz, J. (2022). The Futures of Europe: Society 5.0 and Industry 5.0 as Driving Forces of Future Universities. Journal of the Knowledge Economy, 13(4), 3445-3471. https://doi.org/10.1007/s13132-021-00854-2

Dalkir, K. (2025). Handbook of inclusive knowledge management. CRC Press, Abingdon, Oxon.

Davenport, T. H. (1994). Saving IT's Soul: Human Centered Information Management. Harvard Business Review, March-April, 72(2), 119-131.

Davenport, T., & Prusak, L. (1998). Working Knowledge: How Organizations Manage What They Know. Ubiquity 2000, August, Article 6 (August 1 - August 31, 2000). https://doi.org/10.1145/347634.348775

de Holan, P. M., & Phillips, N. (2004). Organizational forgetting as strategy. Strategic Organization, 2(4), 423-433. https://doi.org/10.1177/1476127004047620

Du Plessis, M. (2007). The role of knowledge management in innovation. Journal of knowledge management, 11(4), 20-29.

Durst, S. (2024). A plea for responsible and inclusive knowledge management at the world level. VINE Journal of Information and Knowledge Management Systems; 54(1), 211–219. https://doi.org/10.1108/VJIKMS-09-2021-0204

Durst, S., & Foli, S. (2024). Responsible and inclusive knowledge management made concrete. In *Handbook of Inclusive Knowledge Management: Ensuring Inclusivity, Diversity, and Equity in Knowledge Processing Activities* (pp. 1-12). CRC Press.

Durst, S., & Khadir, Y. (2025). Towards Responsible Knowledge Management. In S. Durst and Y. Khadir (eds) Knowledge Management at the Crossroads. Navigating Risks and Benefits (pp. 79-88). Springer, Cham.

Durst, S., & Zieba, M. (2018). Mapping knowledge risks: towards a better understanding of knowledge management. Knowledge Management Research & Practice, 17(1), 1-13, https://doi.org/10.1080/14778238.2018.1538603

Dyllick, T., & Muff, K. (2016). Clarifying the Meaning of Sustainable Business: Introducing a Typology From Business-as-Usual to True Business Sustainability. Organization & Environment, 29(2), 156-174. https://doi.org/10.1177/1086026615575176

Ferreira, J., Mueller, J., & Papa, A. (2020). Strategic knowledge management: theory, practice and future challenges. Journal of Knowledge Management, 24 (2), 121–126. https://doi.org/10.1108/JKM-07-2018-0461

McGill, M. E., & Slocum, J. W. (1993). Unlearning the organization. Organizational Dynamics, 22(2), 67–79. https://doi.org/10.1016/0090-2616(93)90054-5

Merhi, M. I. (2023). An evaluation of the critical success factors impacting artificial intelligence implementation. International Journal of Information Management, 69, 102545.

Mökander, J., Sheth, M., Gersbro-Sundler, M., Blomgren, P., & Floridi, L. (2022). Challenges and best practices in corporate AI governance: Lessons from the biopharmaceutical industry. Frontiers in Computer Science, 4, 1068361. https://doi.org/10.3389/fcomp.2022.1068361

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041.

Schneider, J., Abraham, R., Meske, C., & Vom Brocke, J. (2023). Artificial Intelligence Governance For Businesses. Information Systems Management, 40(3), 229–249. https://doi.org/10.1080/10580530.2022.2085825

Williams, C., & Durst, S. (2019). Exploring the transition phase in offshore outsourcing: Decision making amidst knowledge at risk. Journal of Business Research, 103, 460-471.

Zack, M.H. (2002). Developing a knowledge strategy. California Management Review, 41(3), 125-223.

Susanne Durst, Dr., is Professor at the Institute for Knowledge and Innovation (IKI-SEA) at Bangkok University and visiting professor at the Department of Business and Economics at Reykjavik University.

Address: IKI-SEA – BU International, Library Building – Room C6-204, 9/1 Moo 5 Phaholyothin Road, Klong 1, Klong Luang, Pathumthani 12120, Thailand, Tel.: +66(0)2-407-3888 ex 2425, E-Mail: susanne.d@bu.ac.th
ORCID: https://orcid.org/0000-0001-8469-2427

Received: 19.12.2025 | Revised: 30.01.2026 | Accepted: 02.02.2026

Intimate Machines, Disturbed Minds: Managing the Affective Cost of AI

Leona Chandra Kruse, Patrick Mikalef

Abstract: As artificial intelligence (AI) systems increasingly simulate empathy and engage relationally with users, they begin to influence the affective fabric of organizations. Drawing on insights from information systems research, organizational science, and management theory, we argue that existing theoretical and governance frameworks are ill-equipped to account for these developments. We distinguish external regulation from internal governance and propose a preventive–corrective framework for addressing affective costs across work contexts. Building on this framework, we outline a research agenda that challenges prevailing assumptions in management and strategy, calling for renewed attention to affect, relational dynamics, and emotional architectures in organizations shaped by affective AI.

Keywords: Affective AI, AI Governance, Human-AI interaction, Responsible AI, Organizational affect

Vertraute Maschinen, verstörte Geister: über die affektiven Kosten von KI und ihre Steuerung

Zusammenfassung: Da künstliche Intelligenzsysteme (KI) zunehmend Empathie simulieren und mit Benutzern in Beziehung treten, beginnen sie, die affektive Struktur von Organisationen zu beeinflussen. Auf der Grundlage von Erkenntnissen aus der Informationssystemforschung, der Organisationswissenschaft und der Managementtheorie argumentieren wir, dass bestehende theoretische und Governance-Rahmenwerke nicht ausreichend sind, um diesen Entwicklungen Rechnung zu tragen. Wir unterscheiden zwischen externer Regulierung und interner Governance und schlagen einen präventiv-korrektiven Rahmen vor, um affektive Kosten in verschiedenen Arbeitskontexten anzugehen. Auf der Grundlage dieses Rahmens skizzieren wir eine Forschungsagenda, die gängige Annahmen in Management und Strategie in Frage stellt und eine erneute Aufmerksamkeit für Affekte, Beziehungsdynamiken und emotionale Architekturen in Organisationen fordert, die von affektiver KI geprägt sind.

Schlüsselwörter: Affektive KI, KI-Governance, Mensch-KI-Interaktion, Verantwortungsvolle KI, Organisatorischer Affekt

The Shift: From Automation to Affective AI

For centuries, society has been concerned about the prospect of automation: intelligent machines might deskill workers or even render human labor obsolete. Today, however, the most consequential shift concerns intelligent machines entering affective domains once reserved for humans (Zao-Sanders, 2025). Contemporary artificial intelligence (AI) systems increasingly operate not as neutral tools but as affective conversational partners, offering reassurance, mentoring, or even companionship—they emerge as *affective AI*. Managers and employees increasingly rely on them for counsel, conflict navigation, and stress management at work (McKinsey, 2025).

Patterns previously observed in adolescents or clinical contexts (e.g., Chandra Kruse, et al., 2023) now echo emerging dynamics in organizational life, where affective AI is woven into everyday work (Langhof & Güldenberg, 2022). A large-scale study documents a marked rise in business use cases related to emotion, coaching, and companionship (Zao-Sanders, 2025). Yet it remains uncertain how these developments will reshape the business environment. As affective AI engages employees and stakeholders in more interactive and resonant ways, organizations face new dynamics of emotional contagion, trust, and collaboration.

If earlier technologies automated the hand and the mind, affective AI reaches for the heart. By simulating warmth and empathy, these systems blur boundaries between professional and personal, task and feeling, instrumentality and intimacy. They generate new forms of attachment and introduce dynamics that existing organizational theories and governance frameworks are poorly equipped to address. The questions are *(1) What happens when machines do not merely execute tasks but enter the affective domain of organizations? (2) How can we study and govern affective AI in organizations?*

The Asymmetry: Homo Sentiens vs. Affectus Simulatus

At the core of contemporary encounters between people and affective AI lies a fundamental asymmetry in what it means to feel. Humans are *homo sentiens*, embodied beings whose perceptions, emotions, and lived experiences shape interpretation, interaction, and action. Managers and employees do not merely process information; organizational life unfolds through affect as much as through rational calculation. Affective AI, by contrast, operate in the domain of *affectus simulatus*.

Drawing on Jean Baudrillard's Simulacra and Simulation (1981), AI-generated empathy is not authentic *affectus* but a simulation detached from any original emotional experience. What appears as warmth, patience, or reassurance is a performance—a sign of feeling without the feeling itself. Yet simulated affects can have real consequences. Once severed from their referents, signs may reorganize reality rather than merely represent it. Affective AI can reshape the emotional architecture in organizations.

Despite this, management and human resource management (HRM) theory offer limited conceptual grounding for these developments. Affective and relational processes are typically treated as outcomes rather than as mechanisms of organizing. Dominant theories—agency theory, transaction cost theory, and human capital theory—privilege incentives, contracts, and monitoring, while largely neglecting emotion, attachment, and the felt dimensions of coordination and collaboration. As a result, the emotional architecture of organizations remains undertheorized.

The Affective Cost: Individuals and Teams in Organizations

Recent studies suggest that AI chatbots are reshaping how employees seek reassurance and feedback in organizational settings, introducing risks of affective dependency and blurred distinctions between relational support and simulated care (Richet, 2025). Another study reports that AI's tendency toward sycophancy distorts collaboration and undermines critical inquiry by privileging emotional validation over accuracy (Naddaf, 2025). Longitudinal studies further show that individuals experiencing anxiety or depression increasingly form affective attachments to AI systems (Huang et al., 2024).

These risks become critical for organizations as affective AI, being pseudo sentient, assumes roles historically tied to the *homo sentiens*. Their presence fosters affective dissonance, alters relational agency, and deteriorates psychological safety. Anthropomorphized interfaces of affective AI risk generating pseudo-relationships that replace human ones and undermining interpersonal trust. Recognizing this affective cost is essential for understanding how individuals and organizations can learn to self-regulate in environments where *affectus simulatus* becomes part of everyday work.

The cost of affective AI in organizations often emerges first at the individual level before diffusing across organizations. Affective costs are not understood here primarily in financial terms. Rather, they also encompass non-monetary dimensions such as individual well-being, perceived strain, productivity, and broader organizational effects. Workplace chatbots that flatter users or align with their assumptions can foster overconfidence and distort judgment. Over time, such altered dispositions propagate through interaction, reshaping how teams develop trust, assign responsibility, and manage vulnerability.

The Rethinking of Governance: Scope and Timing

As affective AI becomes integrated into everyday organizational life, the limits of existing regulation become clear. Frameworks, such as the EU AI Act primarily address technical and ethical risks through transparency, disclosure, and risk classification. Measures such as age assurance (European Parliament, 2025), user notification, and explanation of the artificial nature of AI companions (California Legislature, 2025) mark first attempts to protect users from false intimacy and affective dissonance. Yet these instruments regulate affective AI from the outside. They cannot, on their own, govern how simulated emotion is enacted, interpreted, and absorbed in daily work.

This gap makes governance essential. Governance refers to the internal practices, norms, and structures that shape how people engage with AI in daily work (Papagiannidis et al., 2025). Distinguishing regulation from governance clarifies that affective costs arise both *a priori* and *a posteriori* of human–AI interaction. Preventive approaches seek to limit problematic affective dynamics upstream, while corrective approaches respond to relational and emotional disruptions once AI becomes embedded in work practices. Importantly, many affective AI interactions unfold outside formally governed organizational systems, such as on personal devices or on general-purpose platforms. This fragmentation limits the reach of formal policies and renders governance an uneven context-dependent process rather than a uniformly enforceable one. Together, these approaches show that the affective cost of AI cannot be managed through compliance alone; it requires governance that actively attends to the emotional architecture in organizations. Table 1 summarizes this logic in a 2×2 framework.

		Scope	
		Regulation (External)	**Governance (Internal)**
Timing	**Preventive (Upstream)**	Rules that limit manipulative affective design, require disclosure of simulated emotional cues, and clarify system limitations.	Organizational design choices that promote emotionally responsible AI: moderating anthropomorphic design, calibrating affective responses, and signaling uncertainty.
	Corrective (Downstream)	Oversight, audits, and enforcement mechanisms addressing unintended emotional or relational harms.	Internal processes to monitor AI reliance, manage AI-mediated breakdowns, and reinforce human-to-human interaction where emotional labor is displaced.

Table 1 Preventive and corrective affective mechanisms across regulation and governance

The Research Agenda: Managing the Affective Cost of AI in Organizations

Affective AI does not arrive in a theoretical vacuum. Information systems research, management science, and human-computer interaction have produced rich accounts of how technologies shape work, coordination, and decision-making. Yet these traditions have largely treated technologies as cognitive or structural forces, not as affective participants. As a result, they offer limited guidance for governing systems that simulate empathy, engage relationally, and shape emotional experience. Advancing governance for affective AI therefore requires challenging and extending existing theoretical foundations across four interconnected levels (Table 2).

Layers	Prevailing Assumptions	Affective AI Shift	Governance Focus
AI Agent Layer	Technologies are neutral, technical tools.	AI behaviors acquire emotional meaning and are read as empathy, attention, or personality.	Design and regulate affective affordances to limit manipulative or misleading emotional signals.
Human Layer	Users are rational; deviations stem from cognitive bias.	Affective AI shapes emotion, attachment, and boundary perception.	Develop self-regulation, AI literacy, and affective awareness in human–AI interaction.
Work Layer	Automation improves efficiency without altering emotional labor.	Affective AI redistributes emotional work and reshapes roles and control.	Redesign work and oversight to manage emotional labor and preserve human judgment.
Relational Layer	Coordination is a human–human process.	AI mediates interaction, feedback, and trust in groups.	Build relational infrastructures that integrate AI while sustaining human connection.

Table 2 Our research agenda

At the ***AI agent layer***, information systems research on affordances, sociomateriality, and human–AI collaboration provides a strong starting point for understanding how

AI shapes organizational action. Governing affective AI requires a shift from functional to affective affordances—how behaviors such as mirroring, adaptation, or monitoring acquire emotional meaning and how they are interpreted. At the ***human layer***, work on human–algorithm interaction has largely focused on cognitive bias. Future research can build on organizational behavior work on emotional regulation and mindfulness to examine how employees learn to discern simulated empathy, manage attachment to AI agents, and maintain psychological boundaries in the face of constant affective feedback—and how to develop these skills.

At the ***work layer***, work design and digital transformation research show how technology reshapes roles, tasks, and control. Organizational change research can help design affective readiness for AI: how should roles be redesigned when traditionally human-centered roles are partially- or entirely- taken over by AI agents? What does "human in the loop" mean when the loop is affective as much as cognitive? Finally, at the ***relational layer***, theories of psychological safety, sensemaking, and coordination must be revisited to account for non-human actors that mediate feedback, trust, and interaction in teams and organizations. Research is needed on how governance structures, routines, and norms can cultivate healthy relational infrastructures that preserve human connection while productively integrating AI mediation.

Combined, these four layers frame a research agenda that addresses both the use of affective AI in organizational life and the design of affective AI systems. Advancing governance in this domain requires integrating design-oriented and use-centered research to ensure that *affectus simulatus* does not quietly reshape the emotional architecture in organizations.

The Epilogue: Intimate Machine, Disturbed Minds

The rise of intimate machines signals a reconfiguration of the emotional architecture in organizations. Interactions once grounded in relations between sentient beings are increasingly mediated by systems that simulate empathy without ever feeling it. Regulating the affective cost of AI is therefore not just a technical or legal challenge but a deeply human one, requiring that we preserve our capacity to feel, to discern, and to remain authentic in the presence of *affectus simulatus*.

References

Baudrillard, J. (1981). Simulacra and simulation. University of Michigan Press.

California Legislature. (2025). Senate Bill No. 243: Companion chatbots. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB243

Chandra Kruse, L., Bergener, K., Conboy, K., Lundström, J. E., Maedche, A., Sarker, S., Seeber, I., Stein, A., & Tømte, C. E. (2023). Understanding the Digital Companions of Our Future Generation. Communications of the Association for Information Systems, 52, 465-479.

European Parliament. (2025, October 16). New EU measures needed to make online services safer for minors. https://www.europarl.europa.eu/news/en/press-room/20251013IPR30892/new-eu-measures-needed-to-make-online-services-safer-for-minors

Huang, S., Lai, X., Ke, L., Li, Y., Wang, H., Zhao, X., ... & Wang, Y. (2024). AI technology panic—is AI dependence bad for mental health? A cross-lagged panel model and the mediating roles

of motivations for AI use among adolescents. Psychology Research and Behavior Management, 1087-1102.

Langhof, J. G., & Güldenberg, S. (2022). The rise of the robot servant-leaders? Next generation leadership. The International Journal of Servant-Leadership, 16(1), 381–424.

McKinsey (2025). Superagency in the workplace: Empowering people to unlock AI's full potential (AI in the workplace: A report for 2025). https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work

Naddaf, M. (2025). AI chatbots are sycophants – and it's harming science. Nature, 647, 13-14.

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. The Journal of Strategic Information Systems, 34(2), 101885.

Richet, J. L. (2025). AI companionship or digital entrapment? Investigating the impact of anthropomorphic AI-based chatbots. Journal of Innovation & Knowledge, 10(6), 100835.

Zao-Sanders, M. (2025, April 9) How People Are Really Using Gen AI in 2025, Harvard Business Review. https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

Leona Chandra Kruse, Prof. Dr., is a Professor of Information Systems at the University of Agder (UiA), Norway. Her research focuses on how people use information systems for work and leisure and how the technologies should be designed to provide people with support, pleasure, and security. Her research has been recognized with several awards, including the Young Researcher Award 2024 from Agder Academy of Sciences and Letters and the AIS Senior Scholar's best Information Systems publication in 2019 and 2025.

Address: University of Agder, Faculty of Social Sciences, Department of Information Systems, Gimlemoen 25, 4604 Kristiansand, Norway, Phone: +4738 14 18 27,
E-Mail: leona.chandra@uia.no
ORCID: https://orcid.org/0000-0002-9001-3870

Patrick Mikalef, Prof. Dr., is a Professor of Data Science and Information Systems at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. His research examines digital transformation, artificial intelligence, data-driven decision-making, and responsible AI governance in organizations. He has published extensively in leading journals in information systems and management and is actively involved in research initiatives at the intersection of AI, strategy, and governance.

Address: Norwegian University of Science and Technology (NTNU), Department of Computer Science, Sem Sælands vei 7–9, 7034 Trondheim, Norway, Phone: +47 73558995,
E-Mail: patrick.mikalef@ntnu.no
ORCID: https://orcid.org/0000-0002-6788-2277

Swiss Journal of Business

Published on behalf of the Schweizerische Gesellschaft für Betriebswirtschaft (SGB)

Established 1947 as *Die Unternehmung. Zeitschrift für Betriebswirtschaft und Organisation*

ISSN 2944-3741

Submissions:
Prof. Dr. Stefan Güldenberg
EHL Hospitality Business School
EHL Campus Lausanne
Route de Berne 301
CH-1000 Lausanne 25
E-Mail: sjb@nomos-journals.de
www.sjb.nomos.de

Manuscripts and Other Submissions:
All submissions should be sent to the above-mentioned address. There is no liability for unsolicited manuscripts that are submitted. They can only be returned if return postage is enclosed. Acceptance for publication must be made in text form.

With the acceptance for publication, the author transfers the simple, spatially and temporally unlimited right to reproduce and distribute in physical form, the right of public reproduction and enabling access, the right of inclusion in databases, the right of storage on electronic data carriers and the right of their distribution and reproduction as well as the right of other exploitation in electronic form for the duration of the statutory copyright to Nomos Verlagsgesellschaft mbH & Co. KG. This also includes forms of use that are currently not yet known. This does not affect the author's mandatory right of secondary exploitation as laid down in Section 38 (4) UrhG (German Copyright Act) after 12 months have expired after publication.

A possible Creative Commons license attached to the individual contribution, or the respective issue has priority in case of doubt. For copyright, see also the general notes at www.nomos.de/copyright.

Unsolicited manuscripts – for which no liability is assumed – are considered a publication proposal on the publisher's terms. Only unpublished original work will be accepted. The authors declare that they agree to editing that does not distort the meaning.

Advertisements:
Verlag C.H.Beck GmbH & Co. KG
Anzeigenabteilung
Dr. Jiri Pavelka
Wilhelmstraße 9
D-80801 München

Media-Sales:
Phone: +49-89-38189-687
E-Mail: mediasales@beck.de

Publisher and Overall Responsibility for Production:
Nomos Verlagsgesellschaft mbH & Co. KG
Waldseestr. 3–5
D-76530 Baden-Baden
Phone: +49-7221-2104-0
Fax: +49-7221-2104-899
www.nomos.de

Geschäftsführer/CEO: Thomas Gottlöber
HRA 200026, Mannheim

Sparkasse Baden-Baden Gaggenau
IBAN DE05662500300005002266
(BIC SOLADES1BAD)

Frequency of Publication: Quarterly

Customer Service:
Phone: +49-7221-2104-222
E-Mail: service@nomos.de

Supported by the Swiss Academy
of Humanities and Social Sciences (SAGW)