

# TOWARD A DIGITAL HUMANISM-BASED FRAMEWORK FOR RESPONSIBLE ARTIFICIAL INTELLIGENCE

Fei Sun

Mälardalen University Press Licentiate Theses  
No. 383

**TOWARD A DIGITAL HUMANISM–BASED FRAMEWORK  
FOR RESPONSIBLE ARTIFICIAL INTELLIGENCE**

**Fei Sun**

**2026**



Department of Computer Science & Engineering

Copyright © Fei Sun, 2026

ISBN 978-91-7485-755-9

ISSN 1651-9256

Printed by E-Print AB, Stockholm, Sweden

# Abstract

This licentiate thesis establishes a normative and methodological foundation for operationalising Responsible AI (RAI), grounded in the philosophical commitments of Digital Humanism. Despite the proliferation of AI ethics guidelines across policy, technical research, and industry domains, a persistent and well-documented implementation gap remains between high-level ethical principles and their practical implementation in AI engineering. This gap is structural, arising from institutional separation among policy, technical research, and engineering practice, as well as systematic failures to translate abstract values into actionable engineering processes. The thesis argues that addressing this gap requires three elements: a normative foundation that goes beyond compliance-oriented metrics, a principled method for making value trade-offs explicit and open to deliberation, and a concrete mechanism for integrating ethical reasoning across the AI lifecycle. Drawing on Digital Humanism, axiology, and Multi-Criteria Decision Analysis (MCDA), it develops the Digital Humanism AI Ethics Toolkit. Within this toolkit, the H.E.A.R.T. model functions as a decision-support mechanism embedded across design, feedback, and continuous improvement processes. Rather than treating ethics as an external constraint or post-hoc evaluation layer, the toolkit supports reflective and accountable decision-making within existing engineering and governance workflows. Across the included studies, the thesis connects a structural diagnosis of Responsible AI operationalisation barriers with the development of methodological and engineering support for value-sensitive AI design and governance.



*To my family, for their unwavering support.  
To myself, for perseverance.*



# List of Papers

This thesis encompasses five papers that together establish the conceptual, methodological, and practical foundation of the research. The papers are presented in a logical progression from problem diagnosis and normative grounding to methodological development and toolkit design.

- Paper I.** F. Sun, “Bridging the Principle–Practice Gap in Responsible AI: A Cross-Domain Review,” manuscript under review for ORCAS 2026 — The 1st International Workshop on Over-Reliance on Cognitive AI Systems in Safety-Critical Domains, co-located with SAFECOMP 2026, 2026.
- Paper II.** F. Sun and D. Isovich, “Responsible AI under the Philosophical Framework of Digital Humanism,” *Information Theory and Applications*, vol. 32, no. 4, pp. 346–354, 2025.
- Paper III.** F. Sun, D. Isovich, and G. Dodig-Crnkovic, “Axiology and the Evolution of Ethics in the Age of AI: Integrating Ethical Theories via Multiple-Criteria Decision Analysis,” *Proceedings*, vol. 126, no. 1, Art. 17, 2025.
- Paper IV.** F. Sun, D. Isovich, and G. Dodig-Crnkovic, “Operationalizing Pluralist AI Governance with the Integrated Axiology–MCDA Framework,” extended and substantially revised journal version of Paper III, manuscript under review for the *Philosophies* special issue “The First International Online Conference Special Issue of the Journal *Philosophies*: Intelligent Inquiry into Intelligence,” 2026.
- Paper V.** F. Sun, D. Isovich, G. Dodig-Crnkovic, J. Stier, and N. Xiong, “The Digital Humanism AI Ethics Toolkit: Translating Values into Action for Responsible AI,” conference paper for the Innovation and Technology Management Conference (InnoTech 2026), China, May 22, 2026.

*Reuse of included papers follows the applicable publisher policies and license terms. All included papers have been reformatted to comply with the layout and typographic conventions of this thesis.*

## **The author's contribution to the included publications**

The author's contributions vary across the five papers but follow a coherent progression throughout the research programme. In Paper I, the author was solely responsible for the conception, literature review, cross-domain synthesis, argument development, and manuscript preparation. In Paper II, the author led the conceptual framing, literature review, analysis, and drafting of the article; the co-author contributed through discussion, supervision, critical revision, and refinement of the philosophical argument. In Paper III, the author was the principal contributor to the development of the axiological–MCDA integration, the analytical structure of the paper, and the drafting of the manuscript; the co-authors contributed through conceptual feedback, supervision, discussion, and revision. In Paper IV, the author led the extension of the Paper III framework into journal form and carried out its substantial revision, including the development of the tripartite value structure, the operationalisation logic, and the manuscript preparation; the co-authors contributed to conceptual refinement, supervision, and critical review. In Paper V, the author led the design and articulation of the Digital Humanism AI Ethics Toolkit and drafted the manuscript; the co-authors contributed domain expertise, conceptual feedback, supervision, and critical review of the proposed framework. Across all papers, the author was the main driver of the research design, synthesis, and writing.

# Abbreviations

---

<b>Abbreviation</b>	<b>Full Form</b>
<b>ACM</b>	Association for Computing Machinery
<b>AHP</b>	Analytic Hierarchy Process
<b>AI</b>	Artificial Intelligence
<b>AI RMF</b>	Artificial Intelligence Risk Management Framework
<b>CRISP-ML(Q)</b>	Cross-Industry Standard Process for Machine Learning (with Quality Assurance)
<b>DSR</b>	Design Science Research
<b>ELECTRE</b>	Élimination et Choix Traduisant la Réalité; a family of outranking decision methods
<b>EU</b>	European Union
<b>H.E.A.R.T.</b>	Human Dignity Audit, Ethical Co-Design, Accountability Layering, Reflective Evaluation, Transparency & Traceability
<b>IEEE</b>	Institute of Electrical and Electronics Engineers
<b>MCDA</b>	Multi-Criteria Decision Analysis
<b>ML</b>	Machine Learning
<b>NIST</b>	National Institute of Standards and Technology (USA)
<b>OECD</b>	Organisation for Economic Co-operation and Development
<b>ORQ</b>	Overarching Research Question
<b>PROMETHEE</b>	Preference Ranking Organisation Method for Enrichment Evaluations; a family of outranking decision methods
<b>QA</b>	Quality Assurance
<b>RACI</b>	Responsible, Accountable, Consulted, Informed
<b>RAI</b>	Responsible Artificial Intelligence
<b>RQ</b>	Research Question (RQ1–RQ4 denote the four specific research questions)
<b>SDLC</b>	Software Development Life Cycle
<b>SE4AI</b>	Software Engineering for AI-enabled Systems
<b>UNESCO</b>	United Nations Educational, Scientific and Cultural Organization

---



# Contents

ABSTRACT . . . . .	I
LIST OF PAPERS . . . . .	V
ABBREVIATIONS . . . . .	VII
CONTENTS . . . . .	IX
PART I: THESIS SUMMARY . . . . .	1
1 INTRODUCTION . . . . .	3
1.1 The Promise and the Problem . . . . .	3
1.2 Structural Critiques of Contemporary AI Ethics . . . . .	4
1.3 Research Aim . . . . .	5
1.4 Research Questions . . . . .	6
1.5 Scope and Terminology . . . . .	6
1.6 Thesis Structure . . . . .	8
2 THEORETICAL FOUNDATION . . . . .	9
2.1 Responsible AI: Principles and Theoretical Foundations for Operationalisation . . . . .	9
2.2 Digital Humanism: Normative Foundation and Motivation . . . . .	10
2.2.1 Philosophical Origins . . . . .	10
2.2.2 Why Digital Humanism for Responsible AI . . . . .	11
2.2.3 Critiques of Digital Humanism . . . . .	12
2.3 Axiology: Structuring Value Pluralism . . . . .	13
2.4 Multi-Criteria Decision Analysis: Conceptual Rationale . . . . .	14
3 RESEARCH DESIGN AND METHODOLOGY . . . . .	16
3.1 Research Methodology . . . . .	16
3.2 Research Methods . . . . .	17
3.2.1 Integrative Literature Review . . . . .	17
3.2.2 Conceptual Analysis . . . . .	17

3.2.3	Artefact Design . . . . .	17
3.3	Reflexivity . . . . .	18
3.4	Evaluation . . . . .	18
4	THE THREE OPERATIONALISATION GAPS . . . . .	19
4.1	The Abstraction Gap . . . . .	19
4.2	The Contradiction Gap . . . . .	20
4.3	The Technocentric Gap . . . . .	20
4.4	The Interconnection of Gaps and Requirements for Operationalisation . . . . .	21
5	INTEGRATED FRAMEWORK AND TOOLKIT . . . . .	23
5.1	Framework Architecture . . . . .	23
5.2	The Toolkit as Engineering Mechanism . . . . .	24
5.3	The H.E.A.R.T. Model . . . . .	25
5.4	MCDA : Structured Value Trade-Off Deliberation . . . . .	26
5.5	CRISP-ML(Q) Overlay . . . . .	27
5.6	Positioning in the SE4AI Community . . . . .	28
6	PAPER SYNTHESIS AND CONTRIBUTIONS . . . . .	30
6.1	Chapter Overview . . . . .	30
6.2	Synthesis and Contributions . . . . .	30
6.2.1	Answering the Research Questions . . . . .	30
6.2.2	Contribution Claims . . . . .	31
6.2.3	Positioning: Narrow and Deep vs. Broad Operationalisation . . . . .	32
7	LIMITATIONS AND FUTURE WORK . . . . .	33
7.1	Evaluation of the Licentiate Phase . . . . .	33
7.2	Future Work . . . . .	34
8	CONCLUSION . . . . .	36
	REFERENCES . . . . .	37
	PART II: PAPERS . . . . .	45
	PAPER I . . . . .	47
1.	Introduction . . . . .	49
2.	Background: The Responsible AI Landscape and the Operationalisation Problem . . . . .	50
2.1	The Emergence of Responsible AI . . . . .	50
2.2	The Operationalisation Problem . . . . .	51
3.	Three Perspectives on Responsible AI . . . . .	52

3.1	Policy Perspective . . . . .	54
3.2	Technical Perspective . . . . .	54
3.3	Industry Perspective . . . . .	55
4.	Three Operationalisation Gaps . . . . .	55
4.1	Abstraction Gap . . . . .	56
4.2	Contradiction Gap . . . . .	57
4.3	Technocentric Gap . . . . .	57
5.	Discussion: Implications for Future Operationalisation . . . . .	58
5.1	Why the Gaps Are Structural, Not Incidental . . . . .	58
5.2	Requirements for a Future Operationalisation Approach . . . . .	59
5.3	Limitations of This Review . . . . .	60
6.	Conclusion . . . . .	60
	References . . . . .	61
PAPER II . . . . .		67
1.	Introduction . . . . .	69
2.	Philosophical Principles of Digital Humanism . . . . .	69
2.1	Human-Centred Design and Cultural Sensitivity . . . . .	69
2.2	Democracy, Inclusion, and Ethical Universalism . . . . .	70
2.3	Rational Inquiry and Thoughtful Development . . . . .	70
2.4	Transparency, Accountability, and Ethical Oversight . . . . .	70
2.5	Human Leadership and Social Responsibility . . . . .	71
2.6	Regulation, Education, and Ethical Frameworks . . . . .	71
3.	Why Digital Humanism Matters for AI . . . . .	71
4.	Digital Humanism as a Philosophical Foundation for Responsible AI . . . . .	72
5.	Challenges and Conclusion . . . . .	73
	References . . . . .	74
PAPER III . . . . .		75
1.	Introduction . . . . .	77
2.	Theoretical Foundations . . . . .	78
2.1	Responsible AI: From Principles to Practice . . . . .	78
2.2	Digital Humanism: A Philosophical Foundation . . . . .	78
2.3	Axiology and Ethical Pluralism . . . . .	78
2.4	Multi-Criteria Decision Analysis (MCDA) in Ethical AI . . . . .	79
3.	Integrating the Axiology–MCDA Framework . . . . .	79
3.1	Normative Foundation . . . . .	80
3.2	Value Classification . . . . .	80
3.3	From Values to Action: MCDA Operational Method . . . . .	80
3.4	Ethical Outcome . . . . .	81

4.	Illustrative Scenario: Ethical Evaluation of AI Diagnostics in Healthcare . . . . .	81
4.1	Scenario and System Alternatives . . . . .	82
4.2	Applying the MCDA Framework . . . . .	82
4.3	Insights . . . . .	83
5.	Conclusions . . . . .	83
	References . . . . .	85
PAPER IV . . . . .		89
1.	Introduction . . . . .	91
2.	Philosophical Foundations of Axiology . . . . .	92
2.1	Classical Foundations: Moore and Ross . . . . .	92
2.2	Contemporary Value Pluralism . . . . .	93
2.3	Axiology and AI Ethics . . . . .	94
2.4	A Tripartite Classification of Values for AI Ethics . . . . .	94
2.4.1	Intrinsic Values . . . . .	95
2.4.2	Instrumental Values . . . . .	95
2.4.3	Relational Values . . . . .	95
2.4.4	Advantages of the Tripartite Structure . . . . .	96
3.	Normative Foundations: Responsible AI and Digital Humanism . . . . .	96
3.1	Responsible AI: From Principles to Embedded Ethics . . . . .	96
3.2	Digital Humanism: Technology in Service of Human Flourishing . . . . .	98
3.3	Ethical Pluralism as the Integrating Principle . . . . .	98
4.	The Integrated Axiology–MCDA Framework . . . . .	99
4.1	Architectural Components and Normative Foundations . . . . .	99
4.2	Axiological Adaptation of MCDA . . . . .	100
4.3	Formal Structure . . . . .	100
4.4	Methodological Procedure . . . . .	101
5.	Case Study: AI Diagnostics in Healthcare . . . . .	102
5.1	Scenario Description . . . . .	102
5.2	MCDA Analysis . . . . .	103
5.3	Sensitivity Analysis . . . . .	104
5.4	Insights and Limitations . . . . .	104
6.	Cross-Domain Applications . . . . .	105
6.1	Education . . . . .	105
6.2	Criminal Justice . . . . .	106
6.3	Finance . . . . .	106
7.	Strengths and Limitations . . . . .	107
7.1	Strengths of the Framework . . . . .	107
7.2	Limitations . . . . .	108
7.3	Comparison with Alternative Approaches . . . . .	108

8. Conclusion . . . . .	109
References . . . . .	111
 PAPER V . . . . .	 117
1. Introduction . . . . .	119
2. Dominant Paradigms and Limitations . . . . .	120
2.1 Ethical Values in AI: Foundations and Limits . . . . .	120
2.2 Current State of Responsible AI Tools . . . . .	121
2.3 Systemic Limitations . . . . .	121
3. Theoretical Framework: Digital Humanism . . . . .	122
3.1 Principles of Digital Humanism . . . . .	122
3.2 Digital Humanism as a Foundation for RAI . . . . .	122
4. The Digital Humanism AI Ethics Toolkit . . . . .	123
4.1 Framework Overview . . . . .	123
4.2 Layered Architecture . . . . .	124
4.2.1 Foundational Principles (Why) . . . . .	124
4.2.2 Governance and Oversight (Who) . . . . .	124
4.2.3 Design and Development Tools (How) . . . . .	125
5. Toolkit Methods for Applied AI Ethics . . . . .	126
5.1 H.E.A.R.T. as the Main Toolkit Method . . . . .	126
5.1.1 Applying H.E.A.R.T. in AI Practice . . . . .	127
5.1.2 Engineering Integration with CRISP-ML(Q) . . . . .	127
5.2 MCDA and the Tripartite Value Taxonomy . . . . .	129
6. Positioning the Toolkit in the SE4AI Community . . . . .	130
7. Conclusion . . . . .	131
References . . . . .	133



**Part I:**

---

# **Thesis Summary**



# 1 Introduction

This thesis addresses a structural gap between AI ethics principles and engineering practices. Artificial intelligence systems increasingly shape decisions across many domains. However, existing ethical frameworks have had limited influence on everyday AI development, and ethical guidance remains disconnected from the development lifecycle. This thesis argues that the gap persists not because of a lack of awareness or motivation, but because ethical guidance is difficult to translate into day-to-day engineering practice. To respond, the thesis develops a normative foundation grounded in Digital Humanism, a method for handling value conflicts, and a way to integrate ethical considerations across the AI development lifecycle.

## 1.1 The Promise and the Problem

Over the past decade, ethical guidelines for AI have proliferated. By 2019, a major review catalogued 84 high-level AI ethics guideline documents produced by governments, corporations, and civil society organisations [1]. The values identified across these documents, including fairness, transparency, accountability, privacy, safety, and human autonomy, were broadly consistent across countries and sectors [1]. More recent reviews indicate that the number of such guidelines has since grown substantially [2]. To an outside observer, the ethical governance of AI might therefore appear to be well established. However, empirical studies tell a different story: software engineers, data scientists, and machine learning practitioners consistently report that ethical guidelines are rarely integrated into day-to-day engineering workflows and decision-making processes [3, 4]. These studies point to a persistent gap between high-level ethical principles and their translation into engineering practice. Ethical commitments are widely articulated, but they remain weakly embedded in everyday engineering decisions. Bridging this gap is central to this research because it requires an approach that translates abstract values into actionable engineering processes under real-world constraints.

The gap is structural. Policy, technical research, and industry practice operate as distinct domains, each with its own forms of knowledge, expertise, and priorities concerning responsible artificial intelligence (RAI). Although each domain contributes valuable perspectives, none is fully

aligned with the realities engineers face when translating abstract values into operational systems under time pressure and organisational constraints. This study is motivated by that structural misalignment and by the institutional conditions it creates for responsible AI governance. In response, the research is organised around five sequential objectives: to identify the factors contributing to the gap; to establish a normative foundation grounded in Digital Humanism; to develop an axiological account of ethical value pluralism; to develop a structured method for reasoning about ethical value trade-offs; and to consolidate these contributions into a practical toolkit for engineering practice.

## **1.2 Structural Critiques of Contemporary AI Ethics**

This section engages directly with influential structural critiques of contemporary AI ethics. These critiques move beyond scepticism to identify systemic problems that any credible contribution must confront. Mittelstadt [5] argues that principles alone cannot guarantee ethical AI and warns that AI ethics risks reproducing problems seen in bioethics, where guidelines may serve symbolic compliance rather than practical reform. Similarly, Hagendorff [6], based on a review of twenty-two major AI ethics guidelines, finds that many lack enforcement mechanisms and have limited attention to power, structural inequality, and political economy. Bietti [7] critiques ethics washing and warns that public commitments to ethical AI can function as superficial forms of governance rather than substantive institutional reform. Metcalf et al. [8] shows that institutional ethics mechanisms can become procedural compliance exercises with limited impact on core system design decisions. Raji et al. [9] similarly highlight that identifying and tracing harmful impacts remains difficult in practice when auditing mechanisms are not embedded throughout the development lifecycle.

Collectively, these critiques justify this study by showing that many contemporary AI ethics initiatives remain weakly connected to the actual production of AI systems. They point to structural decoupling between ethical discourse and engineering practice. In such contexts, design and implementation decisions are driven mainly by technical feasibility and organisational priorities, while values, rights, and social impacts often remain implicit. As Winner [10] argues, technological artefacts are not neutral; they can embody political qualities and produce political effects through their design. This decoupling therefore has consequences beyond technical inefficiency: it shifts normative decision-making into opaque engineering processes and undermines individual agency. When engineering choices are made without explicit normative boundaries, AI systems

risk privileging institutional efficiency over fundamental human rights and public accountability.

This thesis takes these critiques as a starting point. Instead of rejecting principle-based ethics, it contends that principles must be operationalised. A review of Responsible AI across policy, technical research, and industry reveals three structurally recurring gaps between principles and practice. Responding to these gaps requires more than a new set of principles. It requires a normative foundation, a method for handling value conflicts, and concrete means of embedding ethical reasoning across the AI lifecycle. The following sections develop that response.

### 1.3 Research Aim

This thesis constitutes the theoretical and methodological phase of the doctoral project and provides the conceptual and philosophical basis for subsequent empirical work. It aims to advance the operationalisation of Responsible AI by developing a normative and methodological framework grounded in Digital Humanism. This contribution takes the form of a conceptual AI Ethics Toolkit that makes value trade-offs explicit, supports human deliberation, and integrates ethical reasoning across the AI development lifecycle. This aim is addressed through five specific objectives:

- To identify and analyse the operationalisation gap between Responsible AI principles and AI engineering practice.
- To interpret Responsible AI within the philosophical framework of Digital Humanism, establishing a normative foundation that prioritises human-centred values.
- To analyse ethical values and value pluralism in AI development through axiology, distinguishing intrinsic, instrumental, and relational values to provide a principled basis for context-sensitive ethical reasoning, including structured sensitivity analysis of value trade-offs.
- To explore the potential of Multi-Criteria Decision Analysis (MCDA) as a structured method for making value trade-offs explicit, transparent, and subject to stakeholder deliberation.
- To develop the Digital Humanism AI Ethics Toolkit as a conceptual engineering mechanism that operationalises these commitments across the AI development lifecycle.

## **1.4 Research Questions**

The following research questions specify the aim. Key terms used in these questions, including operationalisation, Digital Humanism, axiology, MCDA, and toolkit, are defined with precise, consistent meanings in Section 1.5.

Overarching Research Question (ORQ): How can the principle–practice gap in Responsible AI be addressed through a normative and methodological framework for operationalisation across the AI development lifecycle?

RQ1: What structural factors underlie the persistent gap between ethical principles and engineering practice in Responsible AI?

RQ2: Why is Digital Humanism a philosophically appropriate foundation for Responsible AI, and what does it offer that technocentric frameworks lack?

RQ3: How can axiology and MCDA be integrated to support transparent and structured reasoning about value trade-offs in AI development?

RQ4: How can a Digital Humanism-based AI Ethics Toolkit be designed to operationalise ethical commitments across the AI development lifecycle?

The four RQs follow a cumulative and sequential logic. RQ1 identifies the structural sources of the principle–practice gap and specifies what any adequate response must address. RQ2 establishes the normative foundation for meeting those requirements. RQ3 operationalises that foundation into a structured method for reasoning about value trade-offs. RQ4 translates these normative and methodological contributions into the design of the Digital Humanism AI Ethics Toolkit. The ORQ spans all four and asks how this progression from diagnosis to foundation, method, and integration forms a coherent framework for operationalising Responsible AI across the development lifecycle.

## **1.5 Scope and Terminology**

The following key terms are used with specific and consistent meanings throughout this thesis.

**Table 1.1: Key terms and definitions used throughout the thesis.**

<b>Term</b>	<b>Definition as used in this thesis</b>
<b>Ethics</b>	The systematic study of values, duties, and principles governing human action. In AI contexts, this includes questions about which values should be embedded in systems and how value conflicts should be resolved [5, 11].
<b>Responsible AI (RAI)</b>	The development and deployment of AI that integrates ethical, legal, and social considerations through structured processes embedded in the development lifecycle [12, 13].
<b>Framework</b>	A structured conceptual architecture that organises principles, methods, and practices into a coherent whole. It is distinct from a theory, which explains phenomena, and a toolkit, which provides operational instruments [14].
<b>Theoretical Foundation</b>	The normative and philosophical premises that orient the framework. It is a normative stance grounding design choices rather than a predictive theory.
<b>Toolkit</b>	A set of operational instruments, such as templates, protocols, checklists, and decision methods, used to implement the framework’s commitments at specific stages of the AI lifecycle [15].
<b>Tooling</b>	The broader ecosystem of software-based tools used in AI development practice. The toolkit proposed here is intended to complement this wider tooling landscape rather than replace it [13, 16].
<b>Operationalisation</b>	The systematic translation of abstract ethical principles into actionable engineering practices, decision criteria, and governance mechanisms [17, 18].
<b>Digital Humanism</b>	A normative framework for digital technologies that holds they must uphold human dignity, autonomy, and democratic participation, and must be designed as sociotechnical artefacts within institutional and cultural contexts. In this thesis, it is adapted as the normative foundation for Responsible AI [19–21].
<b>Axiology</b>	The philosophical study of value. In this thesis, axiology is the umbrella value-theory lens; intrinsic, instrumental, and relational values are treated as its three analytic subcategories for AI development [22].
<b>Sensitivity Analysis</b>	A step in MCDA that tests the robustness of rankings by varying value weights and identifying thresholds at which the preferred alternative changes [23].
<b>Constitutive Community Participation</b>	Direct involvement of affected stakeholders in shaping value priorities, evaluation criteria, and acceptable trade-offs in AI design and governance, rather than treating participation as a secondary or merely consultative add-on [24].
<b>MCDA</b>	A structured decision-support method for comparing alternatives across multiple criteria [25].

*The terms above are defined as they are used in this thesis. Although grounded in existing literature, some definitions reflect the author’s interpretation and synthesis for the purposes of this work.*

## **1.6 Thesis Structure**

Chapter 2 develops the theoretical framework by motivating Digital Humanism as the normative foundation and situating it in relation to axiology and MCDA. Chapter 3 outlines the research design and evaluation strategy. Chapter 4 analyses the three structural operationalisation gaps identified in the literature. Chapter 5 presents the integrated framework and the Digital Humanism AI Ethics Toolkit. Chapter 6 synthesises the included papers and their contributions. Chapter 7 discusses the study's limitations and directions for future work.

In the full compilation thesis, these chapters form the introductory and integrative part of the work, followed by the included papers.

## 2 Theoretical Foundation

### 2.1 Responsible AI: Principles and Theoretical Foundations for Operationalisation

Responsible Artificial Intelligence (RAI) has become the primary approach for addressing ethical issues in the design, development, and deployment of AI systems [12, 13]. As a research and practice domain, it draws on normative philosophy, governance studies, and software engineering. Its purpose is to clarify which properties AI systems should have to be ethically acceptable and which processes are needed to achieve and maintain those properties.

The normative content of RAI is structured around a set of core principles that have gained broad recognition across academic, regulatory, and industry contexts. Floridi et al. [11] proposed an influential synthesis of five principles, namely beneficence, non-maleficence, autonomy, justice, and explicability. As discussed in Chapter 1, reviews of AI ethics guidelines show substantial cross-institutional convergence around values such as fairness, transparency, accountability, privacy, and human oversight [1, 2]. This convergence matters here because it shows that RAI begins from a broadly shared normative core, even if operationalising that core remains difficult in practice.

These commitments have been embedded through a layered governance architecture. The OECD AI Principles [26] and the UNESCO Recommendation on the Ethics of Artificial Intelligence [27] have established international normative reference points. The EU AI Act [28] translates these commitments into binding regulatory obligations through a risk-based classification system. The NIST AI Risk Management Framework [29] and IEEE 7000-2021 [30] provide technical standards intended to support implementation. In parallel, major technology companies have introduced internal governance frameworks to translate Responsible AI principles into development and deployment practices, including Google’s AI Principles [31], IBM’s Principles for Trust and Transparency [32], and Microsoft’s Responsible AI Standard [33]. Together, these policies, standards, and industry initiatives form the governance baseline for the operationalisation developed in this thesis.

For RAI principles to guide practice meaningfully, three conditions must be met. First, effective operationalisation requires a normative foundation that provides stable value commitments and resists reduction to

compliance metrics. Second, it requires a principled method for navigating value conflicts, which are structurally unavoidable in complex socio-technical systems [17, 34]. Third, it requires an engineering mechanism that integrates ethical reasoning into key design decisions throughout the AI development lifecycle [18]. The literature suggests that current RAI frameworks often remain limited across these dimensions. This gap between normative consensus and practical application therefore constitutes the core theoretical and methodological problem addressed in this thesis. The following sections develop the normative foundation, methodological approach, and engineering mechanism through which this thesis addresses these three requirements.

## **2.2 Digital Humanism: Normative Foundation and Motivation**

### **2.2.1 Philosophical Origins**

RAI requires a normative foundation that can provide stable value commitments. This section responds to that requirement by introducing Digital Humanism as the normative foundation of the thesis.

Digital Humanism emerges at the intersection of Enlightenment philosophy, computer science, and social theory [19–21]. Its central statement, the Vienna Manifesto on Digital Humanism [2019], calls for digital technologies to support human dignity, autonomy, and democratic participation rather than subordinate these values to economic efficiency or technical optimisation. Drawing on Enlightenment commitments to rational agency, individual freedom, and moral self-governance, subsequent philosophical works [20, 21] argue that these principles must be critically reapplied to modern digital societies, where algorithmic systems, platform governance, and data-driven infrastructures increasingly challenge democratic self-determination.

Digital Humanism presents two core rejections. It rejects the “mechanistic view” that treats humans as computational units, arguing that human behaviour can be fully represented in data and optimised through algorithms. It also rejects the “animistic view” that attributes human-like moral agency to AI systems and shifts human responsibility onto machines. Instead, Digital Humanism holds that AI systems are sociotechnical artefacts shaped by institutional choices, cultural assumptions, and power relations. This aligns with Coeckelbergh’s relational ethics, which understands ethical meaning as emerging from relationships among humans, technologies, and institutions [35]. From this perspective, responsibility for technological artefacts remains with the human actors and institutions that design, deploy, and govern them. This reinforces the thesis’s broader emphasis on deliberation.

### 2.2.2 Why Digital Humanism for Responsible AI

Digital Humanism establishes normative stability by anchoring its commitments in human dignity, autonomy, and democratic participation as non-negotiable starting points. In doing so, it resists reducing ethical principles to mere performance metrics or compliance thresholds [19, 21]. It addresses value conflicts by conceptualising AI systems as socio-technical artefacts shaped by human decisions, governance arrangements, and competing commitments. This perspective provides the philosophical resources necessary to reason about value trade-offs, rather than obscuring them through purely technical proxies or optimisation criteria [36]. Furthermore, it secures democratic legitimacy by emphasising the right of affected stakeholders to participate in decisions about AI systems that impact their lives. It reinforces this principle through explicit commitments to community engagement and participatory governance [37, 38]. Its growing prominence in interdisciplinary discussions further strengthens its relevance within contemporary governance contexts [39–41].

Nevertheless, three alternative normative traditions merit brief consideration.

The *capability approach* [42, 43] offers a rich account of human flourishing and is sensitive to inequality, but its primary domain is development economics and social policy. It does not offer the institutional grounding or lifecycle-integrated engineering application that RAI operationalisation requires.

*Value-Sensitive Design* [44] offers a valuable set of design-oriented methods and a sustained focus on social context and stakeholder experience. These features are important for the present thesis because they inform the treatment of relational values developed later in the framework. However, as Knobel and Bowker [45] notes, Value-Sensitive Design is primarily oriented toward design-level decisions and does not by itself provide the broader governance structures needed to sustain ethical commitments across organisations and development lifecycles. For this reason, Digital Humanism serves as the overarching normative foundation, while VSD's stakeholder-centred methods inform the toolkit's engagement phase.

*Discourse ethics* [46] specifies procedural ideals for participation but presupposes conditions of communicative equality that are difficult to sustain in contexts of structural power asymmetry [47]. It specifies how decisions should be reached but not what values should guide them [48].

Taken together, these comparisons support the application of Digital Humanism as a normative foundation for Responsible AI across design, deployment, and governance contexts. It thus provides the ethical and in-

stitutional basis for the operational framework developed in the following sections.

### **2.2.3 Critiques of Digital Humanism**

Digital Humanism has been criticised, particularly with respect to cultural scope, conceptions of agency, and conceptual precision.

The Eurocentrism critique. Digital Humanism has been criticised for grounding its claims to universal norms in European Enlightenment values [49, 50]. Birhane [51] shows how individualistic and rationalist assumptions embedded in algorithmic systems can reproduce colonial forms of harm. These harms include extracting data from the Global South without corresponding community benefit and imposing Western classification systems on diverse cultural contexts. This thesis responds by treating cultural pluralism as a core design requirement of RAI. It draws on decolonial AI scholarship [50] and design justice approaches that emphasise community-led participation [52]. Community-led engagement methods are therefore built into the engagement phase of the proposed toolkit to ensure that value identification is not predetermined by the assumptions of researchers or developers.

The anthropocentrism critique. As AI systems become more autonomous and relational accounts of agency gain influence, Digital Humanism's emphasis on human authorship and non-delegable moral responsibility can be questioned from relational perspectives on technology and agency [35, 49]. Critics argue that a strictly anthropocentric view risks overlooking hybrid forms of agency involving both human and non-human actors. This thesis does not claim that the boundary between human and machine agency is fixed. Instead, it treats human moral responsibility as a design commitment rather than a descriptive claim. It also warns that engineering practices that delegate responsibility to algorithms risk encouraging ethical abdication. Accordingly, systems should be designed to preserve meaningful human responsibility, since shifting responsibility to technology may weaken accountability for outcomes.

The vagueness critique. Prem [53] acknowledges that Digital Humanism remains an emerging field and has not yet formed a settled definition. Coeckelbergh [54] similarly argues that its practical and political dimensions require further development. This critique is the most relevant for the present thesis. Axiology and MCDA are introduced precisely to translate the broad commitments of Digital Humanism into context-specific and deliberated value criteria that can guide concrete engineering decisions. Understood in this way, the charge of vagueness becomes a call for operationalisation, which forms the central contribution of this research.

These critiques are not unique to Digital Humanism, but reflect broader tensions in contemporary AI ethics and political philosophy, particularly concerning the universality of normative claims, the distribution of agency in sociotechnical systems, and the translation of abstract values into practice.

### **2.3 Axiology: Structuring Value Pluralism**

Digital Humanism provides a clear normative orientation, and the next step is to determine how these values should be identified, distinguished, and translated into decision-relevant structures for AI development.

Axiology, the philosophical study of value, offers the conceptual foundation for translating normative commitments into decision-relevant structures [22, 55]. In its classical formulation, axiology primarily distinguishes between intrinsic values, which are valuable in themselves, and instrumental values, which are valuable as means to further ends. This distinction has been widely adopted in applied ethics and AI ethics to separate moral constraints from performance and utility considerations.

Even though the distinction between intrinsic and instrumental value is analytically useful, it is insufficient for understanding value phenomena that arise within sociotechnical practice. Many ethically salient values in AI systems cannot be fully captured as either intrinsic ends or instrumental means alone, but rather emerge through relationships among human actors, institutions, and technological artefacts [56, 57]. This is where the category of relational values becomes important.

Relational values were developed most prominently in environmental ethics and sustainability scholarship. They refer to values that arise from relationships, such as care, identity, responsibility, and belonging, rather than values that exist independently or serve only as means to an end. These values are especially important in sociotechnical contexts, where legitimacy, trust, and accountability develop through ongoing interactions among people, institutions, and technologies [57].

For this reason, a relational extension of applied axiology for AI ethics is proposed here using a tripartite framework of intrinsic, instrumental, and relational values. This is a key methodological contribution of the thesis. To our knowledge, existing AI ethics frameworks do not systematically treat relational values as a distinct axiological category within an operational AI ethics framework.

Relational values are introduced to make explicit forms of value that depend on context and interaction and that are often treated as implicit in traditional value frameworks. Within this framework, intrinsic values are treated as non-negotiable moral constraints, such as dignity, jus-

tice, and autonomy. Instrumental values concern means–end performance considerations, including accuracy, efficiency, and reliability. Relational values capture conditions of legitimacy that arise through social, institutional, and human–technology interactions, such as trust, accountability, and meaningful participation [56, 57]. This tripartite structure supports context-sensitive ethical reasoning while preserving analytic clarity for design and governance decisions. In this chapter, axiology is introduced as the conceptual basis for distinguishing value types relevant to Responsible AI. Its methodological role in the research design is developed in Chapter 3, while its operational application within the toolkit is presented in Chapter 5.

## **2.4 Multi-Criteria Decision Analysis: Conceptual Rationale**

The tripartite value structure clarifies what is at stake in AI development decisions, but identifying values alone is not sufficient. Value conflicts are unavoidable in sociotechnical systems, and the challenge is not to remove trade-offs but to address them openly and with clear justification. This section explains the conceptual rationale for using MCDA to connect ethical values to concrete engineering choices. Its full operational workflow is presented in Chapter 5.

MCDA is a family of structured decision-support approaches developed in operations research and management science to evaluate alternatives across multiple and often conflicting criteria [25, 58]. Unlike approaches that reduce decisions to a single metric, MCDA makes explicit the criteria under consideration, their relative importance, and how well each alternative performs across them. Its primary strength is its transparency. By making value assumptions explicit and open to discussion, MCDA helps structure deliberation rather than replace it. The method has already been applied in domains such as environmental policy, healthcare allocation, and infrastructure planning, and its relevance to AI ethics has also been explored in software architecture and algorithmic fairness contexts [59–61]. This track record supports the use of MCDA in the present thesis, which aims to structure ethical evaluation and support transparent and well-justified decision-making. MCDA makes value trade-offs explicit and auditable, supports stakeholder deliberation by rendering the weighting of criteria transparent and contestable, and is flexible enough to accommodate the tripartite axiological structure. This allows intrinsic, instrumental, and relational values to be treated as distinct evaluation criteria.

Sensitivity analysis is an integral part of MCDA as used in this thesis. It examines how evaluation results change when the relative importance of

#### *2.4. Multi-Criteria Decision Analysis: Conceptual Rationale*

criteria varies, making explicit how outcomes depend on underlying value judgements. For example, it can reveal when increasing the weight assigned to accuracy begins to outweigh concerns about privacy or fairness, helping stakeholders reflect on the assumptions behind decisions rather than treating numerical results as definitive.

Importantly, MCDA does not by itself determine the ethically correct decision. Its outputs are decision aids that clarify trade-offs and value dependencies, not definitive answers. This framing preserves the primacy of accountable human judgement and avoids treating numerical results as morally conclusive.

# 3 Research Design and Methodology

## 3.1 Research Methodology

This thesis adopts Design Science Research (DSR) as its overarching research methodology. DSR is a research paradigm developed in information systems and software engineering that treats the design and evaluation of artefacts as a valid and rigorous form of scientific inquiry [15, 62]. Unlike explanatory or predictive research, which focuses on describing or modelling existing phenomena, DSR aims to produce artefacts that address practical problems. Such artefacts may include models, methods, or tools. Their value is assessed primarily in terms of usefulness in the intended context. DSR also requires artefact design to be grounded in a clear problem analysis and supported by an evaluation strategy, even when full empirical evaluation is deferred to a later research phase [14].

DSR is well suited to this thesis because the main research outcome is a design artefact, namely the Digital Humanism AI Ethics Toolkit, rather than a predictive theory or an empirical study. The design is also grounded in a clearly articulated problem analysis. Specifically, it responds to the three structural operationalisation gaps identified in Chapter 4, which satisfies DSR's requirement that artefact design address an evidence-based and well-defined problem. In addition, DSR's distinction between design and evaluation phases aligns well with the licentiate stage. At this stage, the focus is on developing conceptual and methodological foundations, while broader empirical evaluation is planned for the doctoral phase.

DSR also has recognised limitations. Designed artefacts may remain too abstract for practical adoption, and evaluation can be less rigorous than in experimental research designs [14, 63]. These limitations are especially relevant at the licentiate phase, where the toolkit remains primarily conceptual and broader empirical evaluation is still pending. The implications for evaluation and the strategy adopted in this thesis are discussed in Section 3.4.

## **3.2 Research Methods**

### **3.2.1 Integrative Literature Review**

An integrative literature review, following Webster and Watson [64] and Torraco [65], was conducted to map existing RAI frameworks across policy, technical, and industry domains and to identify the structural factors producing the principle–practice gap. Rather than pursuing exhaustive coverage, the review purposively sampled representative, high-quality works across the three domains to identify structural patterns. Works were selected on the basis of relevance, influence, and domain representativeness. Sources included IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar, as well as relevant policy, regulatory, standards, and industry documents published between 2014 and 2025. The review provides the problem diagnosis that grounds the framework developed in the following chapters.

### **3.2.2 Conceptual Analysis**

Conceptual analysis was used to develop the theoretical foundation and to clarify the normative and conceptual requirements that any approach to operationalising Responsible AI must satisfy. Within this thesis, axiology provides the conceptual basis for distinguishing the value categories that inform the toolkit design [22, 55]. MCDA informs how those value categories can be translated into evaluative criteria and structured trade-off deliberation within the artefact [25, 58, 66]. Both are treated here as theoretical building blocks of the artefact rather than as independent research methods. Their conceptual rationale is established in Chapter 2, and their operational application within the toolkit is presented in Chapter 5.

### **3.2.3 Artefact Design**

The toolkit was developed through iterative conceptual design informed by the literature review and the theoretical foundation. The design process translated the three structural operationalisation gaps identified in Chapter 4 into specific toolkit components, each responding to a diagnosed failure in current RAI practice: a lifecycle-integrated ethics mechanism to address the Abstraction Gap; a structured value trade-off method to address the Contradiction Gap; and a workflow overlay to embed both within established machine learning engineering practice to address the Technocentric Gap. The design drew on the methodological foundations established by Belton and Stewart [25] and on preference-learning approaches to weight elicitation described by Hüllermeier and Słowiński [67]. The resulting components are presented and explained in full in Chapter 5.

### **3.3 Reflexivity**

This research is situated within a European academic institution and draws primarily on Western philosophical traditions, which shape its normative orientation. The cultural pluralism critique of Digital Humanism [49, 50] therefore constitutes not only an external challenge but also a limitation of the researcher's own standpoint. Addressing this limitation requires deliberate design choices, such as emphasising community engagement and intercultural value identification rather than assuming cultural neutrality.

The choice of Design Science Research reflects a commitment to producing artefacts that are practically usable in engineering contexts. This orientation has guided the toolkit toward engineering specificity, while potentially underemphasising sociological and political dimensions that are foregrounded more strongly in traditions such as science and technology studies or critical data studies.

The licentiate phase does not include empirical data collection from human participants. All claims are based on conceptual reasoning, literature analysis, and artefact design.

### **3.4 Evaluation**

The toolkit developed in this thesis is a conceptual artefact. Within Design Science Research, evaluation primarily concerns utility and usability [62]. Both require empirical testing in real development environments. This testing is planned for the doctoral phase through naturalistic evaluation, in which the toolkit will be applied and observed in ongoing AI development projects [63]. It will be used under practical engineering conditions and refined iteratively in light of the results.

At the conceptual level, the artefact is designed to address the three operationalisation requirements identified: a stable normative foundation; a principled method for navigating value conflicts; and an engineering mechanism for embedding ethical reasoning into development practice. Its practical utility and usability will remain to be empirically evaluated.

# 4 The Three Operationalisation Gaps

The gap between Responsible AI principles and engineering practice is structural. It reflects an institutional separation between policy, technical research, and industry practice. This chapter presents the main contribution of Paper I: a typology of three recurring structural gaps that help explain why RAI principles often fail to translate into practice. The typology builds on Floridi’s argument that high-level principles are insufficient without concrete mechanisms to guide design and governance decisions [11]. The three gaps are not exhaustive, but they offer a simple way to understand the divide between principles and practice. They highlight problems in translating ethics into engineering practice, managing value conflicts, and avoiding the reduction of ethical issues to technical solutions. Although analytically distinct, the gaps reinforce one another, as discussed below.

## 4.1 The Abstraction Gap

The Abstraction Gap occurs when ethical principles are framed too broadly to guide engineering practice. Concepts such as fairness, transparency, accountability, and autonomy are defined in many different and sometimes conflicting ways across RAI initiatives. In the technical literature alone, more than twenty mathematical definitions of fairness have been identified [68]. In practice, developers are rarely given guidance on which definition is appropriate in a given context.

A further aspect of this gap is the lack of procedural support. Many RAI approaches specify desirable system properties but do not provide step-by-step guidance for implementation. By contrast, software engineering relies on established methods such as architecture evaluation, threat modelling, and test-driven development. RAI lacks comparable procedures for connecting ethical goals to concrete design decisions [18]. As a result, ethical concerns are often addressed after deployment through audits or reviews rather than during the requirements and design phases, where key decisions are made [3].

The EU AI Act illustrates this pattern at the regulatory level. Although it introduces important obligations concerning data governance, risk management, and human oversight, it offers little practical guidance on how

these requirements should be implemented. Responsibility is placed on practitioners without sufficient procedural guidance [69]. This combination of strong normative demands and weak implementation support is characteristic of the Abstraction Gap.

## **4.2 The Contradiction Gap**

The Contradiction Gap arises when RAI approaches promote multiple ethical values that cannot all be satisfied at the same time. This problem is not merely theoretical. The fairness literature shows that calibration, false positive rate parity, and false negative rate parity cannot be achieved simultaneously when base rates differ between groups [70, 71]. As a result, policy documents often endorse goals that are mathematically incompatible in real-world settings.

Conflicts also emerge between different values. Greater transparency about system behaviour can create security risks by making systems easier to probe or attack [72]. Privacy protections can also come into tension with demands for accountability and oversight [73]. Individual autonomy may conflict with collective welfare in public-sector systems that require consistent treatment. These tensions cannot be resolved by adding more principles. Instead, they require explicit processes for weighing trade-offs, involving stakeholders, and documenting decisions.

Industry teams experience this gap directly. They must balance regulatory compliance, technical performance, and social expectations simultaneously, and these sources of normative pressure often point in different directions. As a result, ethical judgement is frequently treated as a compliance task, which limits open discussion of value conflicts and trade-offs.

## **4.3 The Technocentric Gap**

The Abstraction and Contradiction Gaps create pressure to reduce ethical concerns to simpler technical terms. The Technocentric Gap emerges when RAI is treated primarily as a technical optimisation problem. Ethical issues are translated into metrics, benchmarks, and objective functions, often at the cost of their social and political meaning. As Green and Chen [74] show, fairness claims can appear technically neutral even though they depend on prior normative choices embedded in the design and use of risk-assessment systems.

Technical tools such as IBM AI Fairness 360 and the Google What-If Tool can support measurement and model comparison [75, 76]. They are useful for detecting certain forms of disparity and exploring design options. However, these tools embed prior value choices, such as which

#### 4.4. The Interconnection of Gaps and Requirements for Operationalisation

harms are measured, which groups are compared, and which trade-offs are accepted. They offer little guidance on whether these choices reflect an appropriate account of justice in a given context.

The Technocentric Gap also involves limited participation by affected communities in defining standards and evaluation criteria. Datasets, fairness metrics, and explainability methods are often developed without direct involvement from groups most likely to experience harm [24, 50, 77]. This creates both an epistemic limitation, because relevant contextual knowledge is missing, and a democratic failure, because value priorities are set without meaningful stakeholder participation.

#### 4.4 The Interconnection of Gaps and Requirements for Operationalisation

The three gaps are not independent but form a mutually reinforcing structure. When ethical principles are formulated at a high level of abstraction, teams tend either to ignore them or to translate them into measurable proxies, giving rise to the Technocentric Gap. When multiple proxies are applied simultaneously, conflicts become unavoidable, producing the Contradiction Gap. When those conflicts remain unresolved, practice is pushed back toward further abstraction or reliance on implicit, metric-driven decisions. In short, abstraction leads to proxies, proxies generate conflicts, and unresolved conflicts reinforce both abstraction and technocentrism.

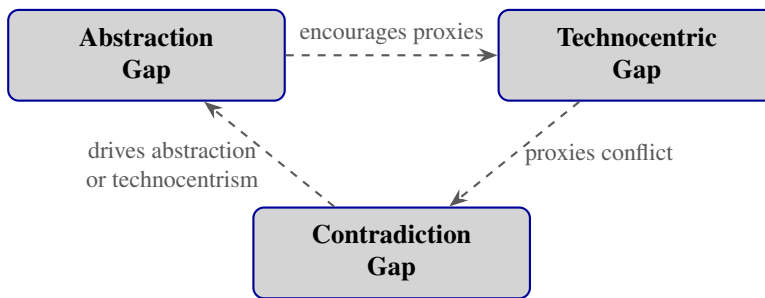


Figure 4.1: The three structural operationalisation gaps and their mutual reinforcement.

These three gaps define the core requirements for Responsible AI operationalisation: lifecycle integration across design, deployment, and post-deployment oversight; structured deliberation over value trade-offs; and meaningful participation by affected communities [78, 79].

These requirements map directly onto the identified gaps and onto the components of the proposed toolkit developed across the thesis and pa-

pers. Lifecycle integration addresses the Abstraction Gap by linking ethical commitments to engineering decisions. Structured trade-off deliberation addresses the Contradiction Gap by making value conflicts explicit and subject to reasoned deliberation. Constitutive community participation addresses the Technocentric Gap by ensuring that value priorities are shaped by those affected [24, 50, 52]. Together, this mapping clarifies the structural logic of the thesis by showing how each design element is intended to address a specific limitation in current RAI practice.

# 5 Integrated Framework and Toolkit

## 5.1 Framework Architecture

This chapter presents the Digital Humanism AI Ethics Framework and Toolkit. The framework refers to the overall architectural and normative structure, while the toolkit denotes the concrete methods and artefacts used to operationalise it. Together, they integrate the normative foundation (Chapter 2), the research design (Chapter 3), and the problem diagnosis (Chapter 4) into a coherent engineering mechanism. In this sense, the framework does not introduce a separate basis for evaluation, but operationalises the normative foundations developed earlier through Digital Humanism, axiology, and MCDA.

The framework is organised into three integrated layers. The term “layer” is used deliberately, since the framework does not proceed through sequential phases. Instead, the layers operate concurrently at different levels of abstraction, each enabling and constraining the others.

The Foundational Principles layer (Why) defines the normative commitments of Digital Humanism, including human dignity, cultural pluralism, democratic control, and transparency, which orient design and evaluation. The Governance and Oversight layer (Who) specifies institutional mechanisms such as ethics boards, audit trails, and appeal mechanisms to ensure accountability and continuity over time. The Design and Development Tools layer (How) provides practical instruments, including the H.E.A.R.T. model, MCDA, bias risk assessment, and human oversight checklists, for embedding these commitments into engineering decisions.

Each layer responds primarily to one of the structural gaps identified. The Foundational Principles layer addresses the Abstraction Gap by establishing a stable normative orientation for interpreting ethical commitments in design and evaluation. The Governance and Oversight layer addresses the Contradiction Gap by providing mechanisms for making value trade-offs explicit, reviewable, and accountable. The Design and Development Tools layer addresses the Technocentric Gap by embedding ethical reflection in concrete engineering practices rather than reducing ethics to technical metrics alone.

Figure 5.1 illustrates how the three layers operate concurrently and reinforce one another. Governance mechanisms enable the accountable use

of design tools; design tools operationalise foundational principles; and foundational principles provide the normative justification for governance requirements. The H.E.A.R.T. model spans all three layers and functions as the lifecycle-integrated ethics mechanism of the toolkit.

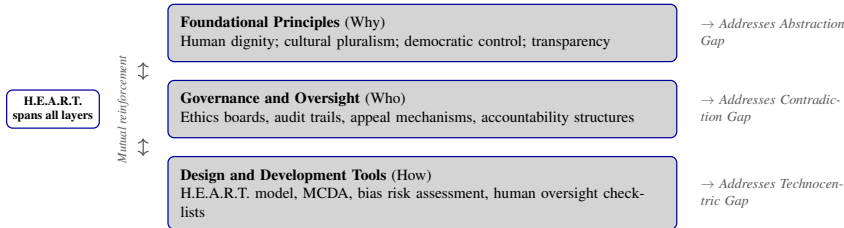


Figure 5.1: Three-layer framework architecture.

The following section explains how this architectural structure is realised in practice through concrete artefacts and decision-support mechanisms.

## 5.2 The Toolkit as Engineering Mechanism

The Digital Humanism AI Ethics Toolkit is a decision-support mechanism composed of the three-layer architecture, four lifecycle phases (engagement, design and development, deployment, and monitoring), and five applied ethics domains operationalised through the H.E.A.R.T. model. Each layer produces concrete outputs that support engineering decisions and provide documentation for governance and review.

The Foundational Principles layer produces a *Value Map* for each project: a concise document that maps the key values relevant to the system and its context. This document identifies intrinsic, instrumental, and relational values relevant to the system under development and situates them in their sociotechnical context. Including relational values brings conditions of legitimacy, trust, accountability, and participation into view alongside rights-based and performance-related concerns.

The Governance and Oversight layer produces *audit-ready documentation*, including ethics board reviews, post-deployment audits, contestability mechanisms, and model cards. These materials provide traceability across the lifecycle and support internal and external oversight, including regulatory assessment under instruments such as the EU AI Act.

The Design and Development Tools layer produces *annotated decision logs* throughout the lifecycle. These logs record ethical assessments, MCDA results, sensitivity analyses, accountability structures, and override decisions.

### 5.3 The H.E.A.R.T. Model

Within this architecture, the H.E.A.R.T. model serves as the toolkit’s primary practical method for operationalising ethical reasoning. As an original contribution of this thesis, it organises ethical reasoning across five domains: Human Dignity Audit, Ethical Co-Design, Accountability Layering, Reflective Evaluation, and Transparency & Traceability. The Reflective Evaluation domain (R) is supported by the MCDA workflow described in the following section, ensuring that value trade-offs are not only deliberated but documented and revisable across the lifecycle. Table 5.1 summarises the model as it functions within the broader framework.

**Table 5.1: The H.E.A.R.T. model and its domains, tools, and outputs.**

	<b>Domain</b>	<b>Core Question</b>	<b>Primary Tools</b>	<b>Output Artefact</b>
<b>H</b>	Human Dignity Audit	Who might this system marginalise or harm?	Empathy mapping; harm scenarios; affected-community consultation	Harm register; stakeholder impact matrix
<b>E</b>	Ethical Co-Design	Have affected groups been involved?	Participatory workshops; surveys; co-creation sessions	Participation log; community requirements document
<b>A</b>	Accountability Layering	Who is responsible at each stage?	RACI matrices; ethics role charts	Accountability map; escalation protocol
<b>R</b>	Reflective Evaluation	Are we doing the right thing, not just the efficient thing?	Retrospectives; ethical red-teaming; MCDA scoring; sensitivity analysis of value weights	MCDA record; sensitivity report; decision rationale log
<b>T</b>	Transparency & Traceability	Can users understand, challenge, or audit behaviour?	Model cards; dashboards; audit logs; contestability channels	Model card; audit trail; contestability specification

The H.E.A.R.T. model is applied iteratively across the AI lifecycle, from early design reviews through deployment readiness assessments to post-deployment monitoring, unlike compliance checklists that are applied only at deployment. In this way, it embeds ethical reasoning as a continuous engineering activity and ensures that value assumptions and trade-offs remain visible, contestable, and revisable over time.

## 5.4 MCDA : Structured Value Trade-Off Deliberation

Chapter 2 established MCDA as a conceptual decision-support approach suited to making value trade-offs explicit, transparent, and subject to stakeholder deliberation. Within the toolkit, MCDA functions as a supplementary deliberative method within the Reflective Evaluation domain of the H.E.A.R.T. model. This section presents the operational workflow for applying MCDA when the H.E.A.R.T. process surfaces value conflicts or requires structured comparison among competing priorities.

Sensitivity analysis is treated as an integral stage of the process rather than an optional add-on. By systematically varying the relative weighting of value criteria, it becomes possible to examine how robust a design decision is under different value priorities and to identify the points at which one ethical consideration begins to outweigh another. This approach makes the assumptions embedded in engineering decisions visible and open to discussion, supporting reflective judgment rather than purely calculative reasoning.

The process follows seven iterative stages:

1. **Context specification.** Define the decision setting, including system boundaries, stakeholders, and the AI design alternatives under consideration.
2. **Criteria identification.** Translate intrinsic, instrumental, and relational values into explicit evaluation criteria relevant to the context, drawing on the Value Map produced in the Foundational Principles layer.
3. **Criterion-based evaluation.** Assess how well each alternative performs on each criterion using empirical evidence, expert judgement, or scenario-based reasoning.
4. **Weight elicitation.** Facilitate stakeholder deliberation on the relative importance of each criterion and assign weights accordingly. This makes explicit how value priorities are determined and supports transparent justification of trade-offs [24, 25, 58].
5. **Scoring and ranking.** Aggregate the performance scores using a weighted-sum model:

$$S(A_i) = \sum_{j=1}^n w_j \cdot x_{ij}, \quad \sum_{j=1}^n w_j = 1, \quad w_j \geq 0.$$

Here,  $S(A_i)$  denotes the overall score of alternative  $A_i$ ,  $w_j$  the weight assigned to criterion  $j$ , and  $x_{ij}$  the performance of alternative  $i$  on

criterion  $j$ . This aggregation is performed only after excluding alternatives that violate intrinsic moral constraints.

6. **Sensitivity analysis.** Systematically vary weight configurations to assess the robustness of the resulting ranking and identify tipping points at which ethical preferences shift. For example, the point at which increasing the weight on accuracy begins to outweigh concerns about fairness or privacy. The outputs of this stage, including the sensitivity report and decision rationale log, feed directly into the Reflective Evaluation domain of the H.E.A.R.T. model.
7. **Reasoned judgement.** Interpret the results collectively, document the ethical rationale as part of an audit trail, and revisit earlier assumptions if contextual or stakeholder conditions change.

Within this framework, MCDA supports ethical deliberation without claiming to determine ethical correctness. Results are understood as decision aids that clarify trade-offs and value dependencies, not as definitive answers. This framing avoids treating numerical results as morally conclusive and reinforces the importance of accountable human judgement.

A recognised limitation of MCDA is the risk of strategic scoring, in which participants may adjust weights or evaluations to favour preferred outcomes [25, 58]. Within this framework, that risk is mitigated through facilitated deliberation led by an ethics steward, explicit documentation of the reasoning behind assigned scores and weights, and calibration exercises based on shared reference cases before evaluating the system under review [24, 25, 58].

## 5.5 CRISP-ML(Q) Overlay

Ethical methods must fit existing engineering workflows rather than create parallel processes that increase effort without clear value. To this end, the toolkit is mapped onto CRISP-ML(Q), a process model for machine learning development with quality assurance [80], linking each phase to a toolkit layer, a H.E.A.R.T. domain, and a defined ethical checkpoint. Table 5.2 shows how the toolkit integrates the normative and methodological foundations into an established ML development workflow. This integration embeds ethical reasoning into established ML engineering practice rather than adding it as a supplementary compliance step.

Table 5.2: CRISP-ML(Q) overlay.

CRISP-ML(Q) Phase	Toolkit Layer	H.E.A.R.T.	Ethical Checkpoint / Output
Business & Data Understanding	Foundational → Value Map	H: Human Dignity Audit	Value Map approved; harm register completed; affected communities identified
Data Engineering	Design Tools → Bias Risk Assessment	E: Ethical Co-Design	Bias risk report; cultural fit assessment; data representativeness score
ML Model Engineering	Design Tools → Human Oversight Checklist	A: Accountability Layering	RACI chart; override documentation; accountability map signed off
Quality Assurance	Design Tools → Ethical QA + MCDA	R: Reflective Evaluation	MCDA score meets threshold; sensitivity analysis report completed; decision rationale documented; red-team findings resolved
Model Deployment	Design Tools → Participatory Deployment Canvas	T: Transparency & Traceability	Model card published; explanation dashboard deployed; contestability specification completed
Monitoring & Maintenance	Governance → Post-deployment Audit	R + T	Audit log current; performance drift report; community feedback reviewed; ethics board notified

## 5.6 Positioning in the SE4AI Community

Within the broader thesis, this framework contributes to Software Engineering for AI-enabled Systems (SE4AI) [16, 81]. SE4AI research has advanced lifecycle models, testing practices, and quality attributes for ML systems, but it often treats ethical concerns as technical properties to be measured or optimised [82, 83]. This framing limits attention to the normative justification, governance, and accountability of ethical decisions.

This thesis extends SE4AI by grounding ethical requirements in Digital Humanism, providing a normative orientation that resists reduction to metrics. It further develops the tripartite axiology–MCDA method as a structured approach to deliberating about value conflicts at concrete decision points in the development workflow. It also embeds ethical reasoning directly into the CRISP-ML(Q) lifecycle through explicit ethical checkpoints and documentation practices. In this way, ethical reasoning is posi-

### *5.6. Positioning in the SE4AI Community*

tioned as a constitutive part of SE4AI practice rather than a supplementary compliance activity.

# 6 Paper Synthesis and Contributions

## 6.1 Chapter Overview

The included papers together constitute the cumulative argument of this thesis. Papers I–V move from problem diagnosis to normative grounding, methodological development, and practical operationalisation. Paper I identifies the structural sources of the disconnect between Responsible AI principles and engineering practice. Paper II establishes Digital Humanism as the thesis’s normative foundation. Papers III and IV develop the axiological and decision-analytic method for structured ethical trade-off reasoning. Paper V consolidates these elements into the Digital Humanism AI Ethics Toolkit as a lifecycle-oriented engineering mechanism.

## 6.2 Synthesis and Contributions

### 6.2.1 Answering the Research Questions

This subsection presents the contributions of Papers I–V by answering each research question in turn. Each research question is restated and linked explicitly to the paper or papers that address it.

RQ1: What structural factors underlie the persistent gap between ethical principles and engineering practice in Responsible AI?

This question is addressed primarily in *Paper I*. The analysis shows that the underlying disconnect is structural rather than epistemic. It arises from the institutional separation between policy, technical research, and industry practice. Paper I identifies three recurring operationalisation gaps: the Abstraction Gap, the Contradiction Gap, and the Technocentric Gap. It then derives three corresponding requirements for operationalisation: lifecycle integration, structured trade-off deliberation, and constitutive community participation by affected stakeholders.

RQ2: Why is Digital Humanism a philosophically appropriate foundation for Responsible AI, and what does it offer that technocentric frameworks lack?

This question is addressed in *Paper II*. Building on the structural diagnosis of RQ1, the paper argues that Digital Humanism provides a normative foundation capable of addressing all three gaps simultaneously. It offers stable value commitments, supports deliberation about value conflicts, and resists the reduction of ethics to technical optimisation. Compared with alternatives, Digital Humanism combines philosophical coherence with practical relevance for Responsible AI operationalisation.

RQ3: How can axiology and MCDA be integrated to support transparent and structured reasoning about value trade-offs in AI development?

This question is addressed jointly in *Papers III and IV*. Paper III develops the axiological and MCDA foundation for structured trade-off reasoning and value clarification. Paper IV extends this foundation through a value-sensitive, tripartite structure (intrinsic, instrumental, and relational), a seven-stage decision process, and sensitivity analysis. It also shows how structured trade-off reasoning can be documented, reviewed, and embedded into engineering workflows through concrete outputs. These papers provide a principled method for addressing value conflicts that would otherwise remain implicit.

RQ4: How can a Digital Humanism-based AI Ethics Toolkit be designed to operationalise ethical commitments across the AI development lifecycle?

This question is addressed in *Paper V*. The paper presents the Digital Humanism AI Ethics Toolkit as an integrated engineering mechanism that translates the normative and methodological contributions of earlier papers into practice. It introduces a three-layer architecture, the H.E.A.R.T. model, and lifecycle integration points that support implementation, review, and adoption across different development contexts.

ORQ: How can the principle–practice gap in Responsible AI be addressed through a normative and methodological framework for operationalisation across the AI development lifecycle?

Through this cumulative progression, Papers I–V collectively answer the overarching research question. Across the included papers, the argument moves coherently from diagnosis to foundation, method, and implementation, demonstrating how Responsible AI can be operationalised across the development lifecycle.

### **6.2.2 Contribution Claims**

Table 6.1 summarises the core contributions of the thesis and aligns them with the RQ–paper mapping presented in the previous subsection.

**Table 6.1: Core contribution claims and their primary evidence base.**

<b>Contribution</b>	<b>Type</b>	<b>Primary Evidence</b>
A typology of three structural operationalisation gaps (Abstraction, Contradiction, Technocentric) arising from the institutional separation of policy, technical research, and industry practice.	Analytical / theoretical	Paper I
A normative argument for Digital Humanism as the philosophical foundation for Responsible AI.	Conceptual / philosophical	Paper II
An integrated axiology–MCDA approach that establishes relational value as a distinct evaluative contribution and enables structured ethical trade-off reasoning.	Methodological + operational	Papers III and IV
The Digital Humanism AI Ethics Toolkit as a consolidated engineering mechanism integrating a three-layer architecture, the H.E.A.R.T. model, MCDA support, and lifecycle integration.	Design artefact (SE4AI)	Papers IV and V
A unified synthesis connecting diagnosis, normative foundation, method, and toolkit implementation into a coherent pathway for Responsible AI operationalisation.	Synthesis	Papers I–V

### 6.2.3 Positioning: Narrow and Deep vs. Broad Operationalisation

Operationalisation research faces a genuine strategic choice between narrow, domain-specific depth and broad, cross-domain architecture [17, 18, 84]. Both approaches are valuable. Narrow operationalisation can achieve high contextual fidelity by embedding domain-specific regulation, institutional workflows, and stakeholder constellations, particularly important in high-stakes domains such as healthcare, public administration, or criminal justice [18].

This thesis adopts the broader approach. The three gaps identified in Chapter 4 are structural rather than sector-specific: they arise from persistent institutional separations between policy, technical research, and engineering practice. Because the problem is cross-domain, the initial response must also operate at a cross-domain level, establishing normative and methodological infrastructure that enables subsequent narrow-and-deep empirical instantiations in the doctoral phase. The proposed toolkit is general in structure but contextual in use: its MCDA component requires local criteria definition and stakeholder weighting, and the H.E.A.R.T. model is intended to be adapted to domain-specific constraints and governance requirements.

# 7 Limitations and Future Work

## 7.1 Evaluation of the Licentiate Phase

The licentiate phase provides the foundation for the broader research programme. Its main contribution is the development of a coherent normative and methodological framework that connects Digital Humanism, axiology, MCDA, and engineering practice through the proposed toolkit. These contributions are primarily analytical rather than empirically validated, reflecting both the scope and the methodological constraints of this stage.

The work is conceptual in nature and does not involve the collection of empirical data from AI developers, affected communities, or governance practitioners. Claims regarding the toolkit's potential effectiveness therefore derive from the existing literature and the internal coherence of the framework rather than from direct observation of its application in practice. This positioning is consistent with the role of the licentiate phase within Design Science Research, which emphasises establishing a sound theoretical and methodological foundation for later empirical evaluation [14, 62].

The literature review adopts a purposive sampling strategy that prioritises work addressing gaps in the operationalisation of AI ethics. As a result, areas such as legal liability, AI rights, normative political theory, and critical data studies receive comparatively limited attention. In particular, decolonial and critical race perspectives highlight forms of harm that are often insufficiently captured by technical or metric-based approaches [50, 77]. These perspectives deserve more substantive engagement in future work.

The research is conducted primarily within a European institutional and philosophical context. While the framework explicitly endorses cultural pluralism [50, 52], it has not been developed through participatory processes with communities outside this context. This constitutes both a methodological and a normative limitation. The doctoral phase should therefore include empirical research conducted in collaboration with communities beyond the European institutional context.

The proposed MCDA-based ethical evaluation framework is subject to further recognised limitations. MCDA approaches are vulnerable to strategic scoring behaviour [25]. Although the toolkit proposes mitigation strategies such as facilitated scoring, mandatory documentation of

rationales, and calibration exercises, their effectiveness has not yet been empirically tested and remains to be validated in real practitioner contexts.

Additionally, the weighted-sum model employed in the toolkit assumes that values can be compared and aggregated meaningfully. Sensitivity analysis cannot fully address cases in which values are genuinely incommensurable, particularly when fundamental constraints such as human dignity are at stake. In such cases, aggregation may be inappropriate, and constraint- or rights-based approaches may be more suitable. Out-ranking methods such as ELECTRE and PROMETHEE could offer better alternatives when values resist straightforward comparison or aggregation. Developing clearer guidance on the conditions that make aggregation defensible is therefore an important task for the doctoral phase.

Finally, scenario-based testing with calibrated reference cases offers only a partial response to the so-called oracle problem in ethical AI evaluation. Judgements about whether a system satisfies requirements related to human dignity ultimately rely on normative evaluation that cannot be fully specified in advance [85]. Addressing this limitation will likely require more deliberative and participatory evaluation approaches, potentially informed by democratic theory, in future research.

## **7.2 Future Work**

The primary objective of the doctoral phase is to empirically validate and further refine the proposed framework across the full AI lifecycle. The Digital Humanism AI Ethics Toolkit will be applied and observed at key lifecycle stages, including stakeholder engagement, design and development, deployment, and post-deployment monitoring. Studying its use under real engineering and governance conditions will enable assessment of its practical relevance, usability, and robustness, in line with established Design Science Research evaluation strategies [63].

Empirical application across multiple organisational and cultural contexts will also allow the limitations identified during the licentiate phase to be addressed systematically. In particular, deploying the toolkit beyond the European institutional environment will support participatory and cross-cultural validation, enabling stakeholder-led value elicitation to inform, and where necessary revise, relational value weightings [50].

Evidence generated across the AI lifecycle will further support refinement of the MCDA components, especially at decision points where value trade-offs arise and values may prove genuinely incommensurable. In such cases, outranking methods such as ELECTRE and PROMETHEE may provide more appropriate alternatives to aggregation-based approaches. Finally, observing ethical reasoning as it unfolds in practice

will help clarify the limits of scenario-based evaluation and inform the development of more robust deliberative procedures for addressing the oracle problem [85].

## 8 Conclusion

This thesis has argued that the gap between Responsible AI principles and engineering practice is fundamentally structural. The persistence of this gap reflects the fact that policy frameworks, technical research, and industry practice operate with different goals, methods, and incentive structures. As a result, the proliferation of additional principles or metrics alone is insufficient to achieve meaningful operationalisation.

In response, the thesis has articulated a coherent framework built around three core elements: Digital Humanism as the normative foundation for Responsible AI; axiology and MCDA as a deliberative approach for explicit and accountable reasoning about value trade-offs; and the Digital Humanism AI Ethics Toolkit as an engineering mechanism that operationalises these commitments across the AI lifecycle through the H.E.A.R.T. model and its integration with CRISP-ML(Q).

A central contribution of the thesis is the extension of the traditional distinction between intrinsic and instrumental value into a tripartite classification that explicitly incorporates relational values such as trust, accountability, and meaningful participation. These relational values capture key conditions of legitimacy for ethical deliberation in practice. Within this framework, sensitivity analysis is positioned as a standard component of the MCDA process, making explicit the value-weight thresholds at which design choices lead to different ethical outcomes and supporting transparent and auditable governance.

Taken together, these contributions establish a clear conceptual and methodological foundation for operationalising Responsible AI within software engineering practice. Ethics is positioned not as an external compliance requirement, but as an integral dimension of engineering work itself. Since AI systems inevitably embed values, those values should be addressed deliberately, transparently, and with meaningful participation from affected stakeholders.

The scope of this licentiate thesis is theoretical and methodological. Its primary contribution lies in providing a defensible foundation and explicit evaluative criteria for the subsequent empirical research planned in the doctoral phase, where the framework will be tested, refined, and assessed in real-world settings.

## References

- [1] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.
- [2] N. K. Corrêa, C. Galvão, J. W. Santos, C. Del Pino, E. P. Pinto, C. Barbosa, D. Massmann, R. Mambrini, L. Galvão, E. Terem, and N. de Oliveira, “Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance,” *Patterns*, vol. 4, no. 10, p. 100857, 2023, originally available as arXiv:2206.11922.
- [3] B. Rakova, J. Yang, H. Cramer, and R. Chowdhury, “Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–23, Apr. 2021.
- [4] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, “Co-designing checklists to understand organizational challenges and opportunities around fairness in AI,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–14.
- [5] B. Mittelstadt, “Principles alone cannot guarantee ethical AI,” *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019.
- [6] T. Hagendorff, “The ethics of AI ethics: An evaluation of guidelines,” *Minds and Machines*, vol. 30, no. 1, pp. 99–120, 2020.
- [7] E. Bietti, “From ethics washing to ethics bashing,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 210–219.
- [8] J. Metcalf, E. Moss, and d. Boyd, “Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics,” *Social Research: An International Quarterly*, vol. 86, no. 2, pp. 449–476, 2019.
- [9] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, “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*, ser. FAT\* ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 33–44.

- [10] L. Winner, “Do artifacts have politics?” *Daedalus*, vol. 109, no. 1, pp. 121–136, 1980.
- [11] L. Floridi, J. COWLS, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Lütge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, “AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations,” *Minds and Machines*, vol. 28, no. 4, pp. 689–707, Dec. 2018.
- [12] V. Dignum, *Responsible Artificial Intelligence*, 1st ed., ser. Artificial Intelligence: Foundations, Theory, and Algorithms. Cham: Springer, Nov. 2019.
- [13] Q. Lu, L. Zhu, J. Whittle, and X. Xu, *Responsible AI: Best Practices for Creating Trustworthy AI Systems*. Addison-Wesley, Dec. 2023.
- [14] S. Gregor and A. R. Hevner, “Positioning and presenting design science research for maximum impact,” *MIS Quarterly*, vol. 37, no. 2, pp. 337–355, 2013.
- [15] R. J. Wieringa, *Design Science Methodology for Information Systems and Software Engineering*, 1st ed. Berlin, Heidelberg: Springer, Nov. 2014.
- [16] Q. Lu, L. Zhu, X. Xu, J. Whittle, and Z. Xing, “Towards a roadmap on software engineering for responsible AI,” in *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, ser. CAIN ’22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 101–112.
- [17] I. Van de Poel, “Translating values into design requirements,” in *Philosophy and Engineering: Reflections on Practice, Principles and Process*, D. Mitchfelder, N. McCarthy, and D. Goldberg, Eds. Springer, 2013.
- [18] J. Morley, C. C. V. Machado, C. Burr, J. COWLS, I. Joshi, M. Taddeo, and L. Floridi, “The ethics of AI in health care: A mapping review,” *Social Science and Medicine*, vol. 260, 2020.
- [19] “Vienna manifesto on digital humanism,” May 2019, Vienna.
- [20] J. Nida-Rümelin and D. Winter, “Humanism and enlightenment,” in *Introduction to Digital Humanism: A Textbook*, H. Werthner *et al.*, Eds. Springer, 2024, pp. 3–16.
- [21] J. Nida-Rümelin and K. Staudacher, “Philosophical foundations of digital humanism,” in *Introduction to Digital Humanism: A Textbook*, H. Werthner *et al.*, Eds. Springer, 2024, pp. 17–30.
- [22] R. S. Hartman, *The Structure of Value: Foundations of Scientific Axiology*. Carbondale: Southern Illinois University Press, 1967.
- [23] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global Sensitivity Analysis: The Primer*. Chichester: John Wiley & Sons, 2008.

- [24] F. Delgado, S. Yang, M. Madaio, and Q. Yang, “Stakeholder participation in AI: Beyond ‘add diverse stakeholders and stir’,” in *Proceedings of the Workshop on Participatory Approaches to Machine Learning, ICML*, 2021.
- [25] V. Belton and T. J. Stewart, *Multiple Criteria Decision Analysis: An Integrated Approach*. New York, NY: Springer, 2002.
- [26] OECD, “Recommendation of the council on artificial intelligence (OECD/LEGAL/0449),” Organisation for Economic Co-operation and Development, Tech. Rep., 2019.
- [27] UNESCO, “Recommendation on the ethics of artificial intelligence,” United Nations Educational, Scientific and Cultural Organization, Tech. Rep., 2021, adopted November 23, 2021; accessed 2026-04-04.
- [28] European Parliament, “Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (AI act),” Official Journal of the European Union, Tech. Rep., 2024.
- [29] NIST, “Artificial intelligence risk management framework (AI RMF 1.0),” National Institute of Standards and Technology, Tech. Rep., 2023.
- [30] IEEE, “IEEE Std 7000-2021: IEEE Standard Model Process for Addressing Ethical Concerns During System Design,” IEEE, Tech. Rep., 2021, accessed: 2026-04-23.
- [31] Google, “AI at google: Our principles,” Google, Tech. Rep., 2018, accessed: 2026-04-02.
- [32] IBM, “Principles for trust and transparency,” IBM, Tech. Rep., 2018, accessed: 2026-04-14.
- [33] Microsoft, “Microsoft responsible AI standard, v2: General requirements,” Microsoft, Tech. Rep., 2022, accessed: 2026-04-02.
- [34] J. Van den Hoven, “Privacy and the varieties of informational wrongdoing,” in *Responsible Innovation*, B. C. Stahl, Ed. Wiley, 2014.
- [35] M. Coeckelbergh, *AI Ethics*. MIT Press, Apr. 2020.
- [36] A. Bennaceur, C. Ghezzi, J. Kramer, and B. Nuseibeh, “Responsible software engineering: Requirements and goals,” in *Introduction to Digital Humanism: A Textbook*, H. Werthner *et al.*, Eds. Cham: Springer, 2024.
- [37] H. Akkermans, “The social responsibilities of scientists and technologists in the digital age,” in *Introduction to Digital Humanism: A Textbook*, H. Werthner *et al.*, Eds. Cham: Springer, 2024.
- [38] S. Winter, “The road less taken: Pathways to ethical and responsible technologies,” in *Introduction to Digital Humanism: A Textbook*, H. Werthner *et al.*, Eds. Springer, 2024.

- [39] H. Werthner, “Digital transformation, digital humanism: What needs to be done,” in *Introduction to Digital Humanism: A Textbook*, H. Werthner *et al.*, Eds. Springer, 2024.
- [40] H. Werthner, E. Prem, E. A. Lee, and C. Ghezzi, *Perspectives on Digital Humanism*. Springer, 2022.
- [41] H. Werthner, A. Stanger, V. Schiaffonati, P. Knees, L. Hardman, and C. Ghezzi, “Digital humanism: The time is now,” *Computer*, vol. 56, no. 1, pp. 138–142, Jan. 2023.
- [42] A. Sen, *Development as Freedom*. Oxford University Press, 1999.
- [43] M. C. Nussbaum, *Creating Capabilities: The Human Development Approach*. Harvard University Press, 2011.
- [44] B. Friedman and D. G. Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, 2019.
- [45] C. Knobel and G. C. Bowker, “Values in design,” *Commun. ACM*, vol. 54, no. 7, pp. 26–28, Jul. 2011.
- [46] J. Habermas, *Moral Consciousness and Communicative Action*. MIT Press, 1990.
- [47] N. Fraser, “Rethinking the public sphere: A contribution to the critique of actually existing democracy,” *Social Text*, no. 25/26, pp. 56–80, 1990.
- [48] M. Coeckelbergh, *The Political Philosophy of AI: An Introduction*. Polity Press, 2022.
- [49] F. Buongiorno and X. Chiamonte, “Do we really need a “digital humanism”? a critique based on post-human philosophy of technology and socio-legal techniques,” *Journal of Responsible Technology*, vol. 18, p. 100080, 2024.
- [50] S. Mohamed, M.-T. Png, and W. Isaac, “Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence,” *Philosophy and Technology*, vol. 33, no. 4, pp. 659–684, 2020.
- [51] A. Birhane, “Algorithmic injustice: A relational ethics approach,” *Patterns*, vol. 2, no. 2, p. 100205, 2021.
- [52] S. Costanza-Chock, *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press, 2020.
- [53] E. Prem, “Principles of digital humanism: A critical post-humanist view,” *Journal of Responsible Technology*, vol. 17, p. Article 100075, 2024.
- [54] M. Coeckelbergh, “What is digital humanism? a conceptual analysis and an argument for a more critical and political digital (post)humanism,” *Journal of Responsible Technology*, vol. 17, p. 100073, 2024.
- [55] N. Rescher, *Introduction to Value Theory*. Prentice-Hall, 1969.

- [56] K. M. A. Chan, P. Balvanera, K. Benessaiah, M. Chapman, S. Díaz, E. Gómez-Baggethun, R. Gould, N. Hannahs, K. Jax, S. Klain, G. W. Luck, B. Martín-López, B. Muraca, B. Norton, K. Ott, U. Pascual, T. Satterfield, M. Tadaki, J. Taggart, and N. Turner, “Why protect nature? rethinking values and the environment,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 6, pp. 1462–1465, 2016.
- [57] M. Coeckelbergh, “Three challenges for a global AI ethics: towards a more relational normative vision,” *AI and Ethics*, vol. 5, pp. 5527–5533, 2025.
- [58] A. Ishizaka and P. Nemery, *Multi-Criteria Decision Analysis: Methods and Software*. Chichester: Wiley, Jun. 2013.
- [59] G. Dodig-Crnkovic and G. Sapienza, “Ethical aspects of technology in the multi-criteria decision analysis,” in *IACAP 2016 Ferrara Conference*, Ferrara, Italy, Jun. 2016, conference held June 14–17, 2016.
- [60] G. Sapienza, G. Dodig-Crnkovic, and I. Crnkovic, “Inclusion of ethical aspects in multi-criteria decision analysis,” in *2016 1st International Workshop on Decision Making in Software Architecture (MARCH)*, Venice, Italy, 2016, pp. 1–8.
- [61] P. Gongora-Salazar, S. Rocks, P. Fahr, O. Rivero-Arias, and A. Tsiachristas, “The use of multicriteria decision analysis to support decision making in healthcare: An updated systematic literature review,” *Value in Health*, vol. 26, no. 5, pp. 780–790, 2023.
- [62] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design science in information systems research,” *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.
- [63] J. R. Venable, J. Pries-Heje, and R. Baskerville, “Feds: A framework for evaluation in design science research,” *European Journal of Information Systems*, vol. 25, no. 1, pp. 77–89, 2016.
- [64] J. Webster and R. T. Watson, “Analyzing the past to prepare for the future: Writing a literature review,” *MIS Quarterly*, vol. 26, no. 2, pp. xiii–xxiii, 2002.
- [65] R. J. Torraco, “Writing integrative literature reviews: Guidelines and examples,” *Human Resource Development Review*, vol. 4, no. 3, pp. 356–367, 2005.
- [66] R. L. Keeney and H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. Cambridge University Press, 1993.
- [67] E. Hüllermeier and R. Słowiński, “Preference learning and multiple criteria decision aiding: Differences, commonalities, and synergies—part ii,” *4OR-Q J Oper Res*, vol. 22, pp. 313–349, 2024.
- [68] S. Verma and J. Rubin, “Fairness definitions explained,” in *Proceedings of the International Workshop on Software Fairness*. New

- York, NY, USA: Association for Computing Machinery, 2018, pp. 1–7.
- [69] M. Veale and F. Z. Borgesius, “Demystifying the draft EU artificial intelligence act,” *Computer Law Review International*, vol. 22, no. 4, pp. 97–112, 2021, preprint DOI: 10.31235/osf.io/38p5f.
- [70] A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.
- [71] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” in *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017.
- [72] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, and A. Madry, “On evaluating adversarial robustness,” 2019, arXiv:1902.06705.
- [73] C. Dwork and D. K. Mulligan, “It’s not privacy, and it’s not fair,” *Stanford Law Review Online*, vol. 66, Sep. 2013, symposium: Privacy and Big Data.
- [74] B. Green and Y. Chen, “Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments,” in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT\*)*, 2019, pp. 90–99.
- [75] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, K. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, “AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias,” *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.
- [76] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. B. Viégas, and J. Wilson, “The what-if tool: Interactive probing of machine learning models,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 56–65, 2020.
- [77] R. Benjamin, *Race After Technology*, 1st ed. Wiley, 2019.
- [78] T. Holstein, G. Dodig Crnkovic, and P. Pelliccione, “Steps towards real-world ethics for self-driving cars: Beyond the trolley problem,” in *Machine Law, Ethics, and Morality in the Age of Artificial Intelligence*, S. J. Thompson, Ed. IGI Global, 2021, pp. 85–107.
- [79] M. Sadek, E. Kallina, T. Bohné, C. Mougénot, R. A. Calvo, and S. Cave, “Challenges of responsible AI in practice: scoping review and recommended actions,” *AI & Society*, vol. 40, no. 1, pp. 199–215, 2025.
- [80] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, and K.-R. Müller, “Towards CRISP-ML(Q): A machine learn-

- ing process model with quality assurance methodology,” *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 392–413, 2021.
- [81] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, “Software engineering for machine learning: A case study,” in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, Montreal, QC, Canada, 2019, pp. 291–300.
- [82] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, and D. Dennison, “Hidden technical debt in machine learning systems,” in *Advances in Neural Information Processing Systems (NIPS)*, 2015, pp. 2494–2502.
- [83] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Douglas, and C. Sanderson, “Software engineering for responsible AI: An empirical study and operationalised patterns,” in *2022 IEEE/ACM 44th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, Pittsburgh, PA, USA, 2022, pp. 241–242.
- [84] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, “The role and limits of principles in AI ethics: Towards a focus on tensions,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 195–200.
- [85] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” 2017.