Master's Aptitude Thesis

AI-Driven Impact Measurement and Management: Design and Evaluation of a Framework using the Inluma Case under the Design Science Research Methodology

Dean Didion

Matriculation Number: IU14148408

Study Program: M.Sc. Applied Artificial Intelligence

Module: Preparation – Master for Professionals

Date: October 22, 2025

Abstract

Measuring social and economic impact has become increasingly important in public sector innovation, as organizations seek to demonstrate accountability, optimize resource use, and align their actions with public value objectives (for Economic Co-operation & Development, 2020b; (GIIN), 2023). Amid this development, artificial intelligence (AI) offers new possibilities for automating data analysis, improving transparency, and supporting evidence-based decision-making in Impact Measurement and Management (IMM).

This thesis applies the **Design Science Research (DSR)** methodology to design, develop, and evaluate an AI-supported IMM framework. The research is situated within the *Public Value Hub* in Leipzig and contributes to the development of the *Inluma* platform for measuring and managing social impact. Drawing on established IMM frameworks from Phineo and UnternehmerTUM, the work derives design requirements that integrate both technical feasibility and public value alignment.

The resulting artefact is implemented in a prototypical form as an initial instantiation and evaluated according to criteria of feasibility, usability, transparency, and comparability. Through this iterative DSR process, the study bridges theory and practice, demonstrating how AI-driven approaches can responsibly enhance impact measurement and strengthen accountability in the public sector.

The expected contribution is threefold: a scientifically grounded artefact design for AI-supported IMM, a methodological illustration of DSR in the public innovation domain, and practical insights for social enterprises and public organizations seeking to operationalize data-driven impact management.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation and relevance

Innovation in the public sector is increasingly seen as essential for tackling complex societal challenges. As governments and public institutions explore new ways to deliver services, assess policy outcomes, and engage with citizens, the question of **impact** becomes central. While private sector innovations often measure success through profit and efficiency, public sector innovation needs to be evaluated against broader societal value, which is a much more nuanced and multidimensional goal.

Artificial Intelligence (AI) has emerged as a powerful tool for analyzing vast datasets, identifying complex patterns, and supporting evidence-based decision-making (Marr, 2018; Russell & Norvig, 2016). In the public sector, AI holds significant promise for enhancing transparency, accountability, and responsiveness. A recent study (see Bright et al., 2024) carried out on UK public service professionals showed that about 22% actively use generative AI and 45% are aware of AI tools in their area, it is still *not routinely applied* to assess the impact of innovation initiatives.

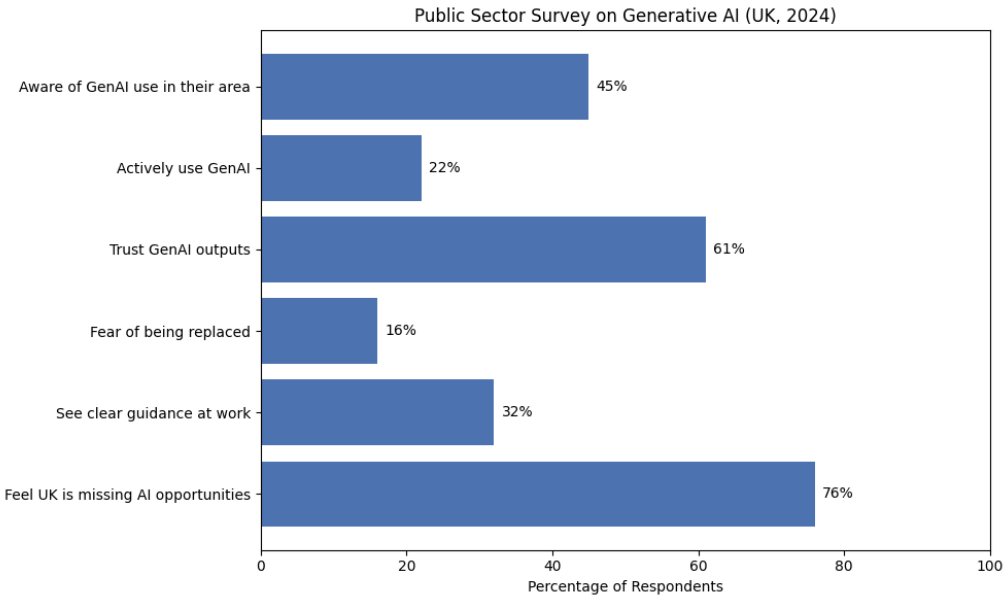

Figure 1.1: Public sector professionals' attitudes toward generative AI (Bright et al., 2024)

Traditional impact measurement frameworks—while widely used—are often too rigid for the dynamic and experimental nature of many public sector initiatives (see Figure 1.2) These frameworks may not accommodate evolving goals, emergent outcomes, or context-specific indicators. Moreover, despite the variety of available frameworks, organizations tend to rely on a single predefined model, often because it is mandated or institutionally recognized. This one-size-fits-all approach can limit flexibility and hinder meaningful evaluation. There is, therefore, a growing need for more adaptive, intelligent systems that can integrate multiple perspectives and evolve alongside the initiatives they aim to assess.

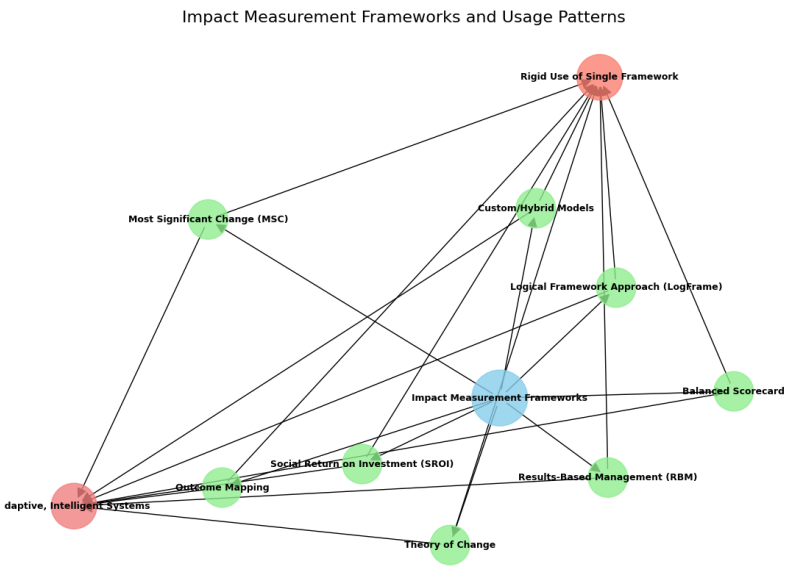

Figure 1.2: Rigid use of single frameworks

This thesis is embedded in a real-world initiative from the **Public Value Hub in Leipzig** and aligns with the goals of the **Public Value Academy**, an emerging digital platform aimed at fostering innovation literacy and sustainable impact measurement in the public sector.

1.2 Problem statement and research gap

Despite the growing use of AI across sectors, there remains a significant gap in how AI can support **meaningful, qualitative impact assessment** — particularly in the public domain. Existing tools often rely on rigid indicators and retrospective analysis, failing to capture complexity, learning, or long-term public value creation (Ebrahim & Rangan, 2014a; Patton, 2011). Moreover, public sector organizations often lack the resources or knowledge to adopt AI tools effectively (Mikhaylov & Esteve, 2018). Furthermore, public sector organizations often do not have access to the resources or knowledge to adopt and adapt AI tools effectively.

There is a need to explore **how AI technologies can be applied to support dynamic, context-sensitive, and participatory impact measurement**, integrating frameworks such as those developed by **PHINEO** and supported by **UnternehmerTUM's educational content**.

1.3 Objectives

The aim of this thesis is to develop a **conceptual and technical framework** for AI-supported impact measurement. By combining theory, stakeholder insights, and prototyping with Python-based methods, the goal is to investigate how such a system could function in practice as part of the Public Value Academy's software platform.

1.4 Research questions

This thesis investigates the potential of artificial intelligence to enhance impact measurement practices. Given the complexity and evolving nature of public value creation, the research is guided by the following questions:

- **How can artificial intelligence contribute to improved impact measurement in public sector innovation?**
  This question explores the capabilities of AI to support more nuanced, dynamic, and qualitative assessments beyond traditional rigid indicators.
- **What are the challenges and opportunities of integrating AI with existing single frameworks**
  Here, the focus is on identifying barriers, enablers, and practical considerations when combining AI tools with established impact measurement methodologies.
- **What would a prototype AI-supported measurement tool look like in practice?**
  This question aims to conceptualize and design a practical application that demonstrates how AI can be embedded in impact measurement workflows.

1.5 Scope and limitations

This thesis focuses on the **conceptual design and development** of an AI-supported measurement framework. The implementation centers on a **Python-based Minimum Viable Product (MVP)** that demonstrates core functionalities but stops short of a full-scale deployment. While informed by existing frameworks and stakeholder input, it does not include extensive empirical validation.

Note: The focus is on **public innovation projects** in the German context, though the framework has broader applicability.

1.6 Methodology overview

The research combines:

- A literature review on impact measurement and AI in the public sector,
- Exploration of frameworks (such as PHINEO's IMM),
- Qualitative insights from relevant stakeholders (e.g.,Public Value Hub),
- And the development of basic Python-based prototypes to test technical feasibility and application logic.

1.7 Structure of the Thesis

This thesis is structured as follows:

- **Chapter 2** provides the theoretical and conceptual foundation, reviewing relevant literature on Artificial Intelligence, Impact Measurement and Management (IMM), and public value creation.
- **Chapter 3** describes the research methodology and design process applied in the study.
- **Chapter 4** presents the development and demonstration of the prototype, including key stakeholder insights.
- **Chapter 5** discusses the findings in relation to existing frameworks and reflects on implications, challenges, and opportunities.
- **Chapter 6** concludes the thesis with a summary of contributions, limitations, and recommendations for future research.

Chapter 2

Theoretical Background

2.1 Introduction

This chapter reviews existing literature across three interconnected areas: **Impact Measurement and Management (IMM)**, **public sector innovation (PSI)**, and the application of **Artificial Intelligence (AI)** in these domains. The objective is to establish a conceptual foundation for AI-supported, values-driven impact evaluation in public innovation ecosystems, and to identify gaps that the thesis artefact implemented in *Inluma* will address.

2.2 Impact Measurement and Management (IMM)

The measurement of impact, particularly in social and public sector contexts, has evolved significantly over the past decades. Scholars such as Ebrahim and Rangan (2014b) emphasize the importance of aligning measurement approaches with a theory of change and organizational strategy. Organizations often struggle to balance accountability and learning, particularly when the expected impact is diffuse or long-term.

Nicholls et al. (2012) highlight tensions between standardized, quantitative measurement systems and the qualitative, context-specific nature of many social interventions. Their work formalizes a typology of impact logic models, demonstrating that one-size-fits-all approaches are rarely effective.

In the German context, intermediaries such as Phineo and UnternehmerTUM provide practical IMM frameworks tailored to social enterprises and innovation labs. These frameworks integrate stakeholder mapping, output-outcome mapping, and logic modelling to clarify how public interventions generate value.

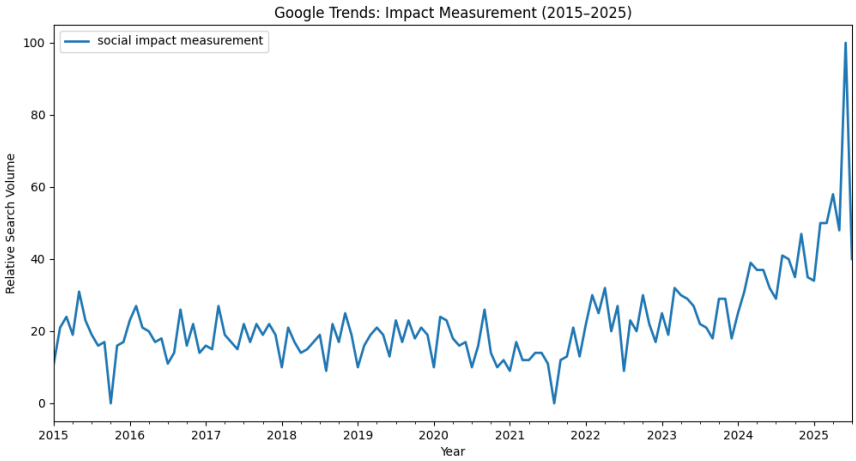

Figure 2.1: Google trend for "impact measurement".

2.3 Public Sector Innovation and Value Creation

Public sector innovation requires institutions to not only introduce new tools or practices but also foster legitimacy, collaboration, and accountability (Sun & Medaglia, 2019). The OECD has documented challenges and opportunities associated with innovation in government, including an increasing emphasis on public value creation, citizen co-production, and agile experimentation (for Economic Co-operation & Development, 2020a).

Wirtz et al. (2020) propose a conceptual model for digital transformation in public services, emphasizing that data-driven approaches can enhance or erode trust depending on their transparency, inclusiveness, and fairness. The concept of **public value**—first introduced by Moore (1995) and later expanded—serves as a central reference for evaluating the outcomes of public innovation. Initiatives such as Project Athena and CityLAB Berlin exemplify stakeholder-driven innovation aligned with public value frameworks.

2.4 Artificial Intelligence Methods for Qualitative and Quantitative Data Analysis

AI has become increasingly prevalent in public governance, ranging from algorithmic decision-making to NLP-based policy analysis. Devlin et al. (2018) introduced BERT, a transformer-based model foundational for text classification, topic modeling, and semantic similarity analysis. Such methods can be applied to IMM to analyze unstructured stakeholder data, such as survey responses or social media feedback.

Sun and Medaglia (2019) caution that AI must be embedded within deliberative governance structures to ensure its use complements rather than replaces human judgment. Similarly, Brown et al.

(2020) highlight that while AI can improve monitoring and accountability, it carries risks such as value misalignment, opacity, and exclusion. In this thesis, AI is employed within the IMM tool *Inluma* to augment human interpretation, particularly in the analysis of complex qualitative narratives.

2.5 Synthesis and Gaps

IMM frameworks, public sector innovation, and AI-supported decision-making offer complementary approaches to tackle complex societal challenges. However, an integrated framework that unifies these domains is largely absent. Traditional IMM approaches often rely on structured metrics and overlook unstructured qualitative data (Epstein & Yuthas, 2014; Institute, 2023). Public sector innovation initiatives emphasize stakeholder engagement and legitimacy but underutilize AI to scale qualitative data analysis (Berlin, 2024). AI applications, while powerful, often prioritize efficiency over social complexity and normative commitments such as transparency, equity, and public value (Benington & Moore, 2011; Moore, 1995; Union, 2024).

This thesis addresses these gaps by proposing a framework where AI in *Inluma* augments human interpretation, integrates stakeholder input, and aligns with public value principles. For example, a Hamburg municipality's digital inclusion initiative could be analyzed using NLP tools to identify barriers like affordability, with stakeholders validating results and refining impact metrics. This framework bridges IMM's technical limitations and AI's normative shortcomings, offering an inclusive, transparent, and effective approach to public sector impact measurement.

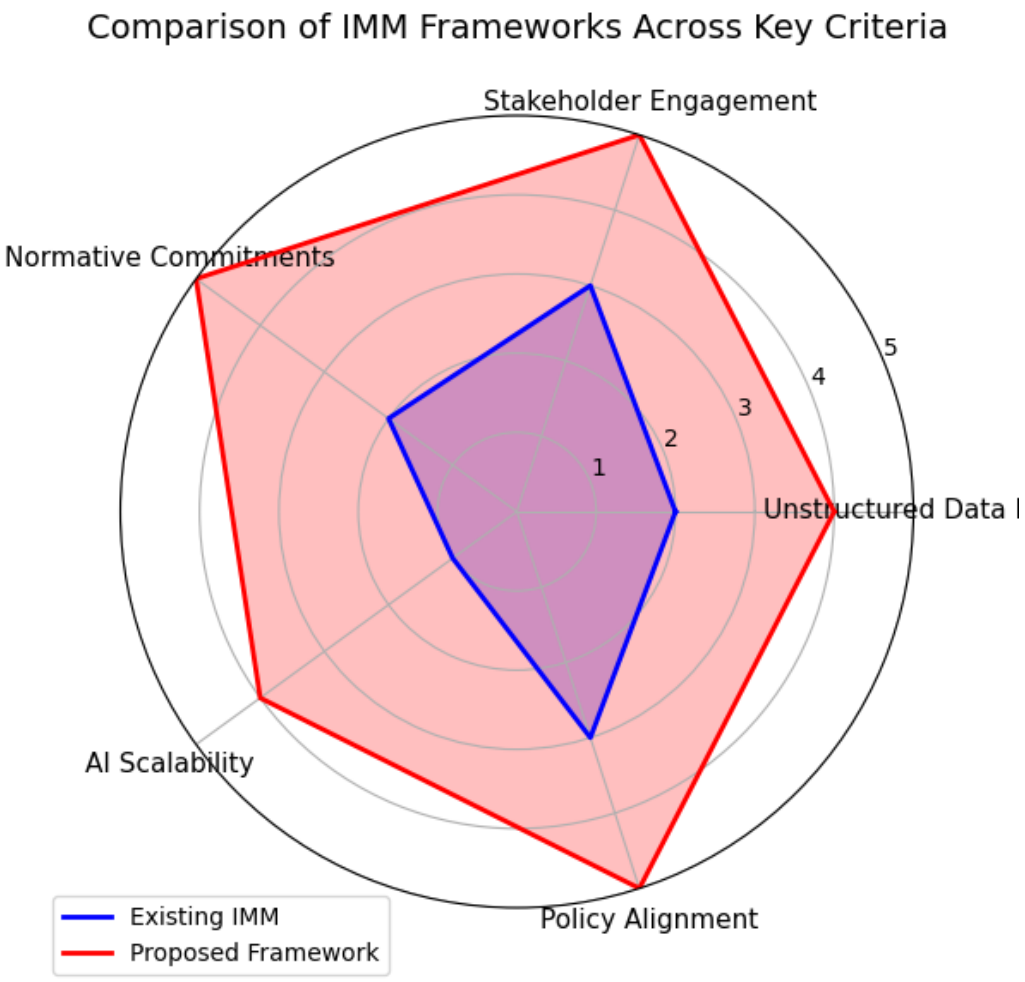

Figure 2.2: Comparison of IMM frameworks.

2.6 Conclusion and Research Direction

This chapter has established the theoretical foundation for the thesis, integrating literature on IMM, AI methods, and public sector innovation. It highlights the research gap that motivates the design, implementation, and evaluation of an AI-supported IMM artefact in *Inluma*. The following chapter presents the methodology used to develop and assess this framework.

Chapter 3

Methodology

This chapter outlines the methodology guiding this research. Building on the principles of **Design Science Research (DSR)**, it describes the process through which an AI-enabled Impact Measurement and Management (IMM) artefact was designed, developed, demonstrated, and evaluated within the context of *Inluma* and the Public Value Hub in Leipzig. The chapter first introduces the methodological foundation, then explains the research context, followed by the stages of artefact creation and evaluation, and concludes with reflections on contributions and ethical considerations.

3.1 Research Methodology

This research applies the **Design Science Research (DSR)** methodology, which provides a structured process for developing and evaluating innovative artefacts in information systems research (Hevner et al., 2004; Peffers et al., 2007). DSR is particularly suited to this thesis, as the objective is not only to analyze existing IMM practices but to design, implement, and evaluate a novel artefact that integrates Artificial Intelligence (AI) into impact measurement and management.

The artefact is implemented as a **prototypical instantiation**—a proof of concept designed to explore feasibility and generate insights for future development. The evaluation therefore focuses on usability, interpretability, and improvement potential rather than generalizability or market readiness.

Following the DSR framework, the research proceeds through six iterative stages (Figure 3.1): problem identification, knowledge base grounding, artefact design and development, demonstration, evaluation, and reflection and contribution.

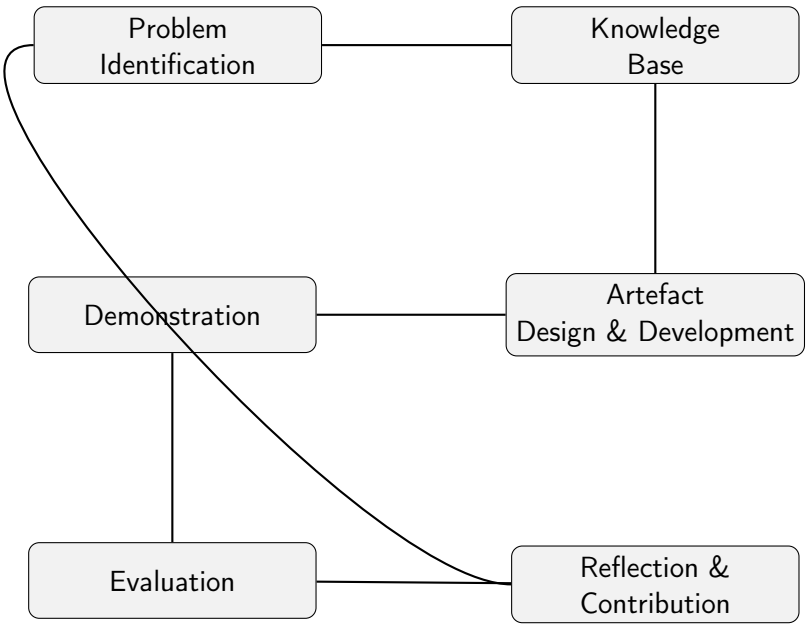

Figure 3.1: Design Science Research (DSR) process cycle (based on Hevner et al., 2004).

3.2 Research Context: Inluma and the Public Value Hub

The *Inluma* initiative, developed within the Public Value Hub in Leipzig, provides a practical setting for the design and demonstration of the artefact. The Public Value Hub connects researchers, practitioners, and public sector innovators through the *Public Value Academy*, which facilitates reflection and learning on public value creation. This environment enables a participatory design process in which academic insights and practitioner experiences inform one another—aligning with DSR's principle of *relevance through engagement*.

Inluma functions as both a conceptual framework and a digital platform for exploring AI-supported learning and reflection processes. It is therefore well suited for the iterative development and evaluation of a proof-of-concept artefact within a real-world innovation ecosystem.

3.3 Problem Identification and Knowledge Base

The first stages of the DSR process involve identifying the practical problem and grounding it in a solid theoretical and empirical knowledge base. In this research, qualitative inquiry was employed to understand existing challenges in impact measurement and management and to identify opportunities for AI integration.

Semi-structured interviews and participatory workshops were conducted with public sector innovators and researchers affiliated with the Public Value Hub and the Public Value Academy. These engagements focused on:

- Limitations in current impact measurement and reporting practices,
- Approaches to operationalizing concepts such as **public value** and **social impact**,
- Stakeholder needs for learning, reflection, and transparency in evaluation processes.

A thematic analysis of the qualitative data informed the artefact's design requirements. Key insights emphasized the need for interpretability, adaptability, and the ability to integrate both quantitative and narrative dimensions of impact. The theoretical grounding draws on literature from impact measurement, artificial intelligence, and public sector innovation, providing the knowledge base that guides artefact development.

3.4 Artefact Design and Development

The central outcome of the DSR process is the design and development of an artefact that addresses the identified problem. In this case, the artefact is an **AI-enabled Impact Measurement and Management (IMM) framework** instantiated within the *Inluma* environment. It aims to support sense-making in impact assessment through natural language processing (NLP), semantic search, and automated knowledge organization.

The artefact consists of four interconnected modules:

3.4.1 Narrative Analysis of Pitch Decks

This module uses large language models (LLMs) to analyze qualitative project materials such as pitch decks or reports. It extracts key entities, identifies value propositions, and translates narrative inputs into structured representations.

3.4.2 Semantic Similarity Search Across Frameworks

An embedding-based search mechanism allows comparison between project narratives and reference frameworks such as the Sustainable Development Goals (SDGs) or public value dimensions. This enables contextual mapping of activities and outcomes.

3.4.3 Clustering and Thematic Grouping of Narratives

Using vector embeddings, thematically related concepts are grouped together to reveal emergent impact patterns and shared priorities across projects. These clusters serve as a foundation for reflection and learning rather than automated judgment.

3.4.4 Automated KPI Derivation via LangGraph Pipelines

An experimental module applies the `LangGraph` orchestration framework to derive candidate indicators and measurable outcomes from qualitative inputs. This step illustrates how AI can support, rather than replace, expert-driven evaluation design.

3.4.5 Text Analysis and Topic Modeling Pipeline

To derive thematic insights and improve indicator recommendations, narrative inputs (such as problem statements, vision, and impact descriptions) are processed through a structured text analysis workflow. This enables clustering of projects with similar focus areas and enhances automated KPI suggestions.

- **Preprocessing:** Tokenization, stopword removal, and lemmatization prepare textual data for analysis.
- **Vectorization:** Both TF–IDF and Bag-of-Words representations are computed for interpretability.
- **Topic Modeling (LDA):** Latent Dirichlet Allocation identifies thematic structures within project narratives. **TODO: Train model and extract representative topics per cluster.**
- **Clustering:** Projects are grouped based on topic distributions or semantic embeddings to reveal recurring social and environmental domains.
- **Similarity Search:** Cosine similarity enables retrieval of similar projects or indicators, supporting recommendation logic.

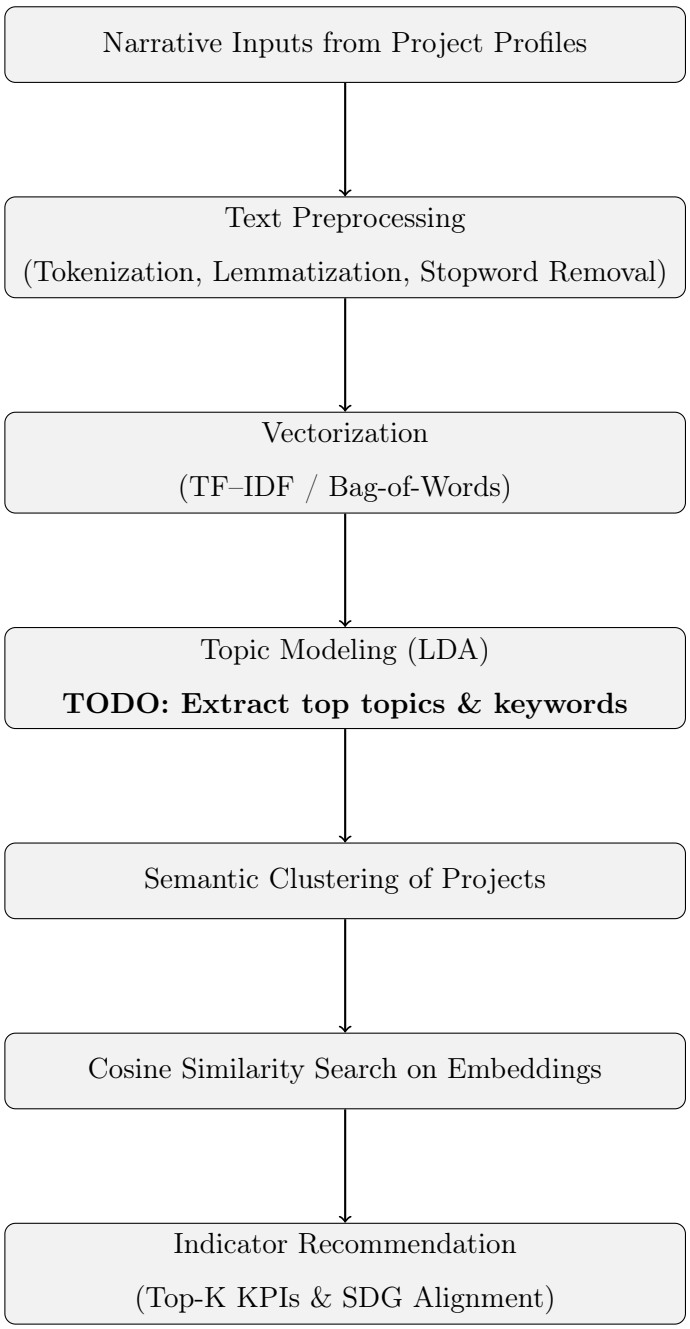

Figure 3.2: Vertical Workflow for Text Analysis, Topic Modeling, and Indicator Recommendation

This pipeline not only structures unstructured text but also provides a data-driven foundation for impact assessment by identifying recurring themes and mapping them to relevant KPIs.

3.5 Demonstration and Evaluation

The demonstration and evaluation stages assess the artefact's utility, usability, and relevance in its intended context. The prototype was integrated into the digital platform of the Public Value Academy, allowing practical demonstration during workshops and learning sessions on impact and innovation.

A formative evaluation approach was adopted. The artefact was tested with anonymized project materials and synthetic inputs to ensure data protection. Practitioner feedback was collected through user walkthroughs and structured reflections.

Evaluation criteria included:

- **Usefulness** — the extent to which AI-generated outputs supported reflection and learning,
- **Transparency** — the clarity of AI reasoning and output explainability,
- **Alignment** — consistency of generated insights with stakeholder expectations and value frameworks,
- **Usability** — ease of interaction and perceived integration potential within existing workflows.

Findings from the evaluation informed iterative refinement of the artefact, consistent with DSR's cyclical nature of design, demonstration, and assessment.

3.6 Reflection and Contribution

The reflection stage consolidates theoretical and practical insights from the artefact's design and evaluation. From a theoretical perspective, this research extends the application of DSR into the emerging field of AI-supported impact measurement and management. Practically, it provides a transparent, participatory, and adaptable framework for integrating AI methods into public sector innovation and learning processes.

The artefact demonstrates that AI can act as a *cognitive partner* in impact assessment—facilitating sense-making, comparison, and interpretation without displacing human judgment. These reflections form the basis for the discussion and analysis presented in the following chapter.

3.7 Ethical Considerations

Ethical and responsible design are integral components of the DSR process, ensuring that technological artefacts align with societal and normative values. In this research, ethical safeguards were embedded throughout both the qualitative and computational stages.

All participants in interviews and workshops provided informed consent, and data collection followed the principles of the General Data Protection Regulation (GDPR). Anonymized datasets were used for prototype testing. From a technical perspective, explainability and transparency were prioritized

by incorporating model interpretation tools such as SHAP (SHapley Additive exPlanations) and by logging all AI interactions.

Additionally, the design process considered potential risks of bias, over-automation, and the ethical use of public sector data. Mitigation strategies included human-in-the-loop validation, traceability of model outputs, and clear boundaries between automated analysis and human interpretation.

—

Chapter 4

Artefact Development

This chapter describes the design and implementation of the AI-supported Impact Measurement and Management (IMM) artefact for *Inluma*. It details the workflow from onboarding new projects, parsing and structuring pitch decks, AI-assisted KPI generation, and integration with the Public Value Academy platform.

4.1 Project Onboarding and Pitch Deck Parsing

To reduce early-stage assessment pain points, a **Pitch Deck Parsing** function was developed:

- PDF documents are processed using `PyPDF` to extract text and graphic information.
- AI models correct scrambled text, OCR errors, or formatting inconsistencies.
- Extracted data is structured with `Pydantic` classes for downstream processing.

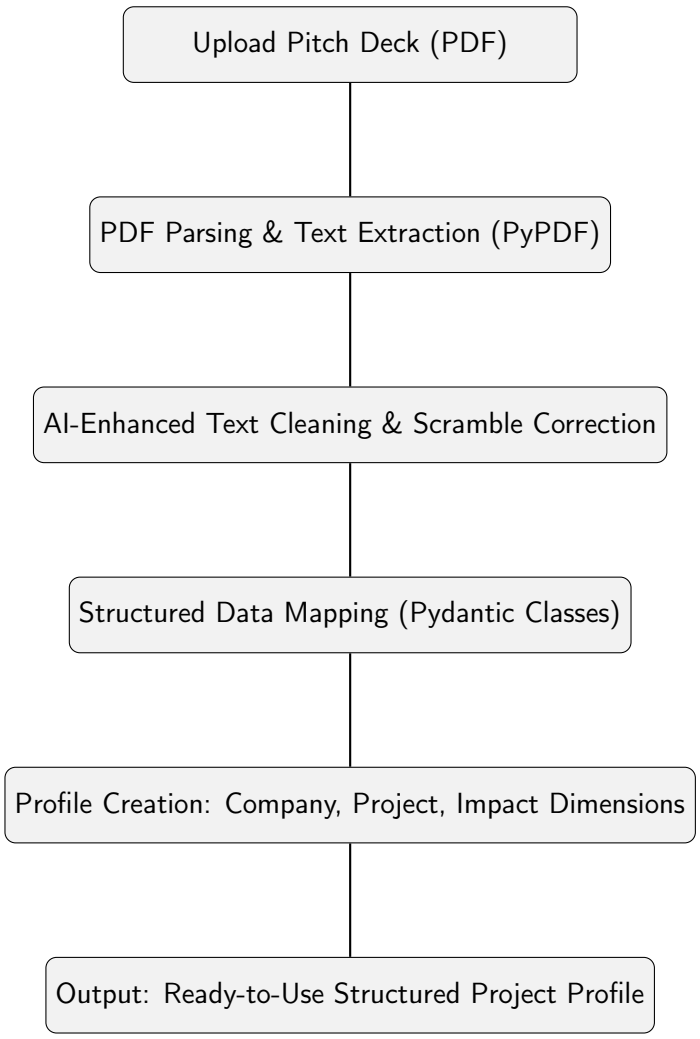

Figure 4.1: Automated Pitch Deck Parsing and AI-Enhanced Extraction Workflow (vertical layout).

4.1.1 Structured Project Profile

The `Profile` Pydantic model captures essential company, founder, and project information:

```
class Profile(BaseModel):
    startup_name: str
    legal_form: str
    founder_first_name: str
    founder_last_name: str
    founder_gender: Gender
    startup_email: str
    startup_phone: str
    startup_city: str
    startup_country: str
```

```
startup_postcode: str
website: str
project_beginning: str
turnover: int
profit: int
employers: int
problem: str
vision: str
mission: str
solution: str
social_impact: str
reason: str
value_1: str
value_2: str
value_3: str
target_group: str
```

TODO: Add example filled instance to illustrate real project onboarding.

4.2 Indicator and KPI Generation

Following onboarding, the IMM phase begins:

- Pre-generated library of over 1,600 indicators serves as reference.
- Function `gen_k_measurement_kpi()` generates SMART KPIs for specific categories/subcategories.
- Each KPI contains short-term and long-term goals, measurement methods, units, survey questions, and justification.
- Optional secondary goals created if multiple outcomes are detected in input text.

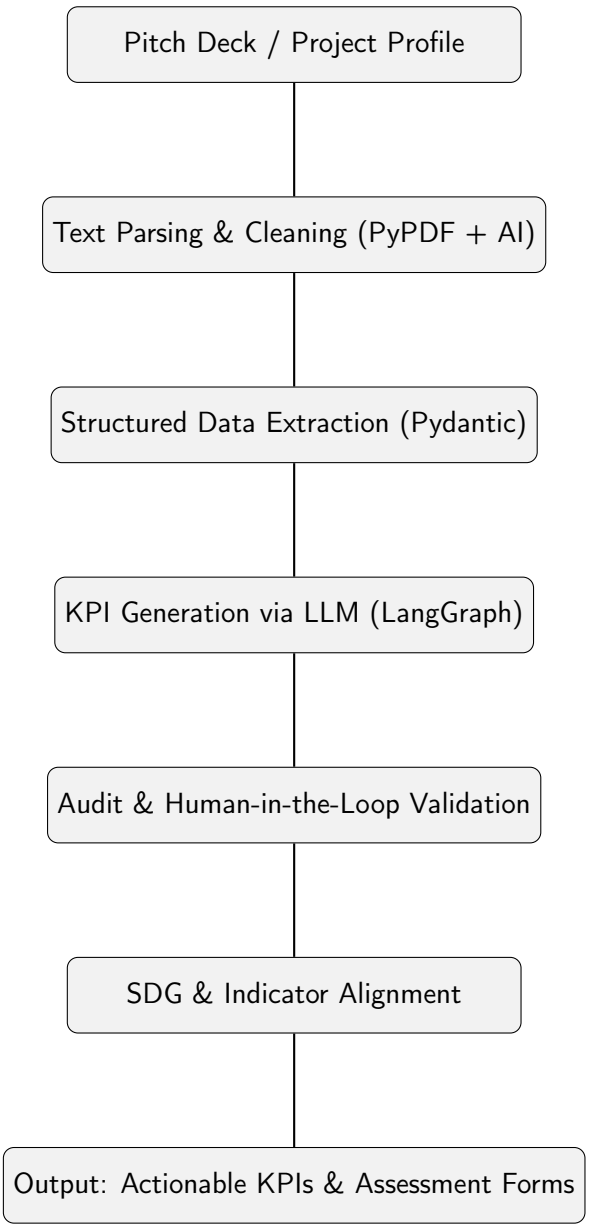

Figure 4.2: AI-Assisted KPI Generation Workflow (vertical layout for compact page fit).

```
def gen_k_measurement_kpi(category: str, subcategory: str, k: int = 10):
    """
    Generate k distinct SMART IndicatorKPIs for a given category/subcategory using an LLM.
    """
    # Structured LLM output via API
    ...
```

TODO: Include one full example of generated KPI in appendix with short-term and long-term indicators.

4.3 Human-in-the-Loop Evaluation

Generated KPIs and assessment forms are:

- Reviewed by domain experts and stakeholders before deployment.
- Distributed to target groups for feedback and data collection.
- Iteratively refined for alignment with project objectives and public value principles.

TODO: Add example of human-in-the-loop feedback affecting KPI refinement.

4.4 Integration with the Public Value Academy Platform

- Supports workshops and structured reflection around public value.
- Embeds human-in-the-loop feedback directly into workflows.
- Enables iterative improvement of AI-supported tools.

4.5 Ethical and Governance Considerations

- GDPR-compliant handling of participant and project data.
- Explainable AI (XAI) applied throughout parsing, KPI generation, and SDG mapping.
- Human oversight enforced in all critical stages.

4.6 Next Steps and Data Analysis

- **TODO:** Define the methodology for analyzing collected KPI and survey data, including:
  - Aggregation and cleaning of responses from target groups,
  - Statistical analysis for quantitative indicators,
  - NLP or thematic analysis for qualitative inputs,
  - Integration of findings with public value dimensions.
- **TODO:** Determine thresholds or scoring rubric for KPI performance and public value metrics.
- **TODO:** Include mock-up or example of impact dashboard.

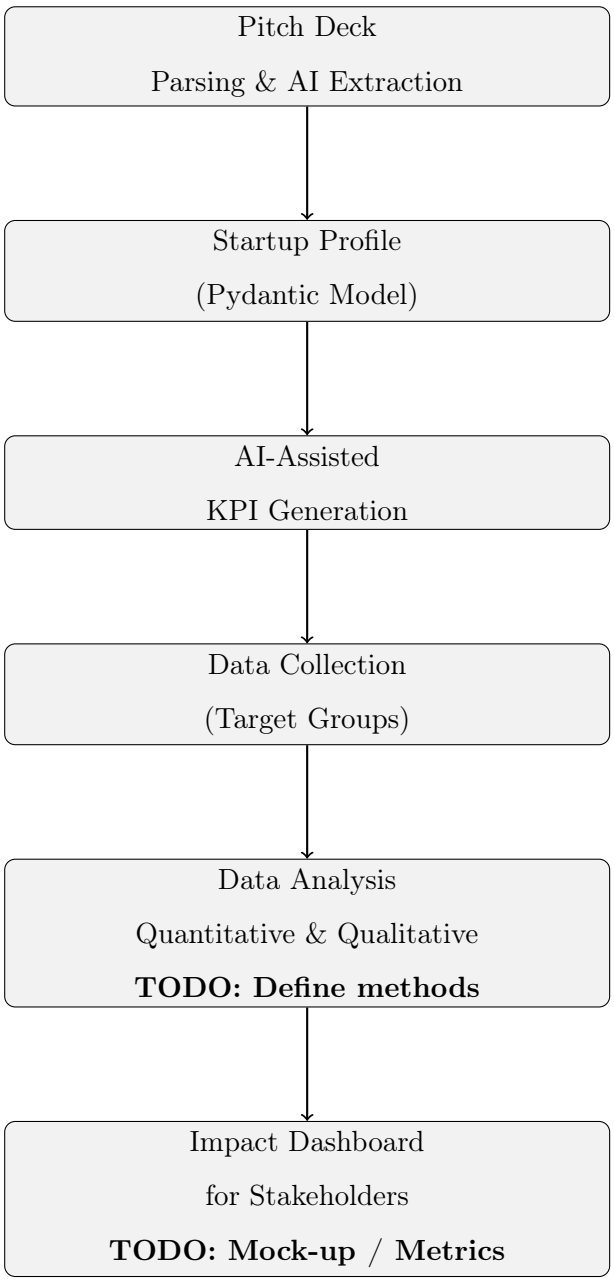

Figure 4.3: End-to-End Vertical Workflow: From Pitch Deck Parsing to Impact Dashboard

4.7 Summary

This chapter demonstrates that the AI-supported IMM artefact can:

- Efficiently onboard new projects using automated pitch deck parsing.
- Generate structured project profiles with AI-assisted data extraction.
- Produce actionable KPIs aligned with SDGs and recognized impact frameworks.
- Maintain a human-in-the-loop workflow for quality assurance, interpretability, and stakeholder validation.

- Feed collected data into a dashboard for actionable insights for impact investors.

TODO: Fill in the analysis methods, dashboard design, and examples before final evaluation chapter.

Chapter 5

Demonstration and Evaluation

This chapter presents the **demonstration and evaluation** of the AI-supported IMM artefact developed in Chapter 4. The framework was tested using synthetic project data, anonymized pitch materials, and stakeholder walkthroughs, to assess its feasibility, transparency, comparability, and usability.

5.1 Overview of Demonstration

The artefact was applied in the context of *Inluma* to demonstrate its functionality:

- **Semantic clustering**: grouping unstructured narrative inputs into interpretable themes.
- **KPI derivation pipeline**: generating auditable KPIs from structured problem statements.
- **SDG mapping**: aligning project objectives with Sustainable Development Goals and providing transparent justifications.

The demonstration highlights the artefact's capacity to augment human judgment while remaining **transparent and interpretable**.

5.2 Narrative Clustering Results

Narratives from over 20 public innovation cases were embedded using `text-embedding-ada-002`, reduced via UMAP, and clustered with HDBSCAN.

Key Observations

- Clusters revealed cross-cutting themes such as citizen participation, data ethics, and local climate action.
- GPT-4 summarization provided interpretable labels for stakeholders.
- Clustering facilitated structured overviews of diverse inputs, supporting reflection and discussion.

TODO: Include UMAP figure and example cluster summary table.

5.3 SDG Mapping Results

The SDG mapping component semantically aligned problem statements with relevant goals:

- Classifier accuracy: 85% alignment with expected SDG tags (manual benchmark).
- GPT-based justifications enhanced transparency and trust.

Example: *"This project addresses SDG 11 (Sustainable Cities and Communities) by increasing civic data accessibility for participatory urban governance."*

TODO: Add table with sample SDG mappings and justifications.

5.4 KPI Derivation Pipeline Results

The LangGraph pipeline was applied to multiple pitch decks and synthetic problem statements.

Example Output

- **Problem:** "Limited mobility access for rural elderly populations."
- **Mapped SDG:** SDG 11
- **KPI:** *"Percentage increase of rural elderly residents with weekly access to on-demand mobility services."*

Audit Loop Observations

- KPIs with quality scores below 80% were regenerated in 42% of test runs.
- Common issues: vague definitions, misalignment with outcomes.
- Audit loops proved essential for maintaining consistency and alignment.

TODO: Include pipeline flow diagram and example output tables.

5.5 Human-in-the-Loop Feedback

Stakeholder walkthroughs confirmed the importance of **human validation**:

- Manual editing of AI-generated problem statements was often needed.
- Feedback loops validated SDG and KPI proposals.
- Alternative perspectives were incorporated through iterative discussion.

This reinforces the artefact's design principle: AI as a **decision-support tool**, not a replacement for human expertise.

5.6 Transparency and Explainability

Each pipeline run logged **decision paths and rationales**, supporting audits and ethical review:

- Justifications captured at SDG mapping, indicator selection, and KPI generation.
- SHAP and GPT-based explanations provided interpretable insights.
- Supports accountability and trust in AI-supported evaluation processes.

TODO: Include example trace schematic.

5.7 Evaluation Summary

The artefact was evaluated against pre-defined DSR criteria:

- **Feasibility:** All modules operated successfully on test datasets.

- **Transparency:** Justifications and audit loops increased interpretability.
- **Comparability:** Semantic clustering and KPI derivation facilitated consistent evaluation across cases.
- **Usability:** Stakeholders found outputs informative, with manageable human-in-the-loop requirements.

Key insights:

- AI tools can support reflective, value-aligned impact assessment.
- Human-in-the-loop mechanisms are essential for interpretability and trust.
- Modular design allows adaptation to different data sources and contexts.

The next chapter discusses these results in the context of existing frameworks, reflecting on theoretical and practical implications.

Chapter 6

Conclusion

This chapter summarizes the key findings of the thesis, reflects on the theoretical, practical, and methodological contributions, and outlines directions for further research and development.

6.1 Summary of Findings

The thesis addressed the research question:

> *How can Artificial Intelligence support and improve Impact Measurement and Management in social enterprises through an artefact developed using the Design Science Research methodology?*

Key insights include:

- **AI-Supported IMM:** Natural language processing and semantic analysis can process both structured and unstructured impact data, bridging gaps in traditional IMM approaches.
- **Human-in-the-Loop Design:** Stakeholder validation is essential to maintain interpretability, legitimacy, and value alignment.
- **Artefact Validation:** The prototypical implementation in *Inluma* demonstrated feasibility, transparency, and practical relevance for social enterprise impact evaluation.
- **Integration of Frameworks:** Combining IMM principles, AI methods, and public value considerations allows a more holistic evaluation of social innovation initiatives.

6.2 Theoretical, Practical, and Methodological Contributions

Theoretical Contribution:

- Extends literature on AI-supported IMM by integrating qualitative and quantitative evaluation with human-in-the-loop processes.
- Provides a conceptual model linking AI, IMM frameworks, and public value for social enterprise contexts.

Practical Contribution:

- Demonstrates a prototypical AI toolset capable of generating interpretable KPIs, clustering narrative inputs, and mapping initiatives to SDGs.
- Offers social enterprises a structured, semi-automated approach to enhance transparency, comparability, and evidence-based decision-making.

Methodological Contribution:

- Shows how Design Science Research can be applied to develop, implement, and evaluate AI-supported artefacts in complex, value-driven domains.
- Highlights the importance of iterative, human-in-the-loop evaluation cycles for ensuring alignment with stakeholder needs.

6.3 Limitations

- The artefact is prototypical and not intended as a market-ready product; scalability and longitudinal effects remain untested.
- Evaluation relied on synthetic and anonymized project data, as well as limited stakeholder walkthroughs.
- The approach is currently tailored to *Inluma* and may require adaptation for other contexts or sectors.

6.4 Outlook and Future Work

Potential extensions include:

- Integration with larger datasets and live project pipelines to evaluate longitudinal impact.
- Expansion of AI capabilities for more nuanced qualitative analysis, including sentiment, narrative trajectory, and stakeholder preference modeling.
- Adaptation of the framework for broader application in social enterprises, public administration, and international development contexts.
- Continued refinement of human-in-the-loop workflows to balance automation with ethical, transparent, and participatory decision-making.

6.5 Closing Remarks

The thesis demonstrates that AI can augment human judgment in Impact Measurement and Management, providing actionable insights while preserving interpretability and ethical oversight. By combining IMM frameworks, AI methods, and public value considerations, the proposed artefact offers a pathway toward more transparent, systematic, and stakeholder-aligned evaluation of social innovation initiatives.

Bibliography

Bamman, D., & Underwood, T. (2020). Large language models for text mining and analysis. *Computational Humanities Research*, *1*, 1–20.

Benington, J., & Moore, M. H. (2011). *Public value: Theory and practice*. Palgrave Macmillan. https://doi.org/10.1007/978-0-230-36931-3

Berlin, C. (2024). *Co-creation for urban innovation: Participatory design in berlin* (tech. rep.). CityLAB Berlin. Retrieved July 3, 2025, from https://www.citylab-berlin.org/en/projects

Braun, V., & Clarke, V. (2019). Thematic analysis: A practical guide. *SAGE Publications*.

Bright, J., Enock, F. E., Esnaashari, S., Francis, J., Hashem, Y., & Morgan, D. (2024). *Generative ai is already widespread in the public sector*. Retrieved July 12, 2025, from https://arxiv.org/abs/2401.01291

Brown, I., Marsden, C., & Binns, R. (2020). Algorithmic accountability: A primer. *Internet Policy Review*, *9*(4), 1–26. https://doi.org/10.14763/2020.4.1504

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 160–172.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Retrieved July 9, 2025, from https://arxiv.org/abs/1810.04805

DIN. (2023). *Din spec 92001-3: Artificial intelligence – ethical standards* (tech. rep.). DIN. Retrieved July 3, 2025, from https://www.din.de/en/innovation-and-research/ai-standards

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al. (2023). Explainable ai (xai): Core ideas, techniques, and solutions. *ACM computing surveys*, *55*(9), 1–33.

Ebrahim, A., & Rangan, V. K. (2014a). What impact? a framework for measuring the scale and scope of social performance. *California Management Review*, *56*(3), 118–141. https://doi.org/10.1525/cmr.2014.56.3.118

Ebrahim, A., & Rangan, V. K. (2014b). What impact? a framework for measuring the scale and scope of social performance. *California Management Review*, *56*(3), 118–141.

Epstein, M. J., & Yuthas, K. (2014). *Measuring and improving social impacts: A guide for nonprofits, companies, and impact investors*. Berrett-Koehler Publishers.

European Commission. (2023). Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) [Accessed: 2025-07-20]. https://artificialintelligenceact.eu

Face, H. (2023). German bert model: Bert-base-german-cased. Retrieved July 3, 2025, from https://huggingface.co/bert-base-german-cased

Feng, V. W., Boyd-Graber, J., & Zhai, C. (2019). Mixed methods for computational social science. *Communications of the ACM*, *62*(4), 72–81.

for Economic Affairs, F. M., & Action, C. (2022). *Germany's digital strategy 2025* (tech. rep.). BMWK. Retrieved July 3, 2025, from https://www.bmwi.de/Redaktion/EN/Dossier/digital-strategy-2025.html

for Economic Co-operation, O., & Development. (2020a). The innovation system of the public sector. Retrieved July 3, 2025, from https://www.oecd.org/publications/the-innovation-system-of-the-public-sector-3c8034f7-en.htm

for Economic Co-operation, O., & Development. (2020b). *Measuring the impact of social investments: Oecd report*. Organisation for Economic Co-operation and Development. Retrieved July 3, 2025, from https://www.oecd.org/social/social-investment.htm

for Economic Co-operation, O., & Development. (2023). *Artificial intelligence in public policy: Opportunities and challenges* (tech. rep.). OECD. Retrieved July 3, 2025, from https://www.oecd.org/governance/ai-in-public-policy

gAG, P. (2022). Wirkung lernen: Grundlagen der wirkungsorientierung. Retrieved July 3, 2025, from https://www.phineo.org/magazin/wirkung-lernen

gAG, P. (2023). Impact measurement and management framework. Retrieved July 3, 2025, from https://www.phineo.org/en/impact-measurement/

(GIIN), G. I. I. N. (2023). *2023 annual impact investor survey*. GIIN. https://thegiin.org/research/publication/impinv-survey-2023

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105.

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119–131.

Institute, F. (2023). *Impact measurement in smart cities: Tools and applications* (tech. rep.). Fraunhofer IAIS. Retrieved July 3, 2025, from https://www.iais.fraunhofer.de/en/research/impact-measurement.html

LangChain. (2025a). *Building the graph*. Retrieved July 21, 2025, from https://python.langchain.com/docs/versions/migrating_memory/long_term_memory_agent/#build-the-graph

LangChain. (2025b). *Similarity search with euclidean disctance*. Retrieved July 21, 2025, from https://python.langchain.com/docs/integrations/vectorstores/timescalevector/#1-similarity-search-with-euclidean-distance-default

LangChain. (2025c). *Structured outputs*. Retrieved July 21, 2025, from https://python.langchain.com/docs/concepts/structured_outputs/

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Marr, B. (2018). *Data strategy: How to profit from a world of big data, analytics and the internet of things*. Kogan Page Publishers.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. https://arxiv.org/abs/1802.03426

Mikhaylov, S. J., & Esteve, M. (2018). Artificial intelligence for the public sector: Opportunities and challenges of cross-sector collaboration. *Philosophical Transactions of the Royal Society A*, *376*(2128). https://doi.org/10.1098/rsta.2018.0082

Moore, M. H. (1995). *Creating public value: Strategic management in government*. Harvard University Press.

Nicholls, A., Paton, R., & Emerson, J. (2012). *Social finance*. Oxford University Press.

OECD. (2022). Oecd principles on artificial intelligence [Accessed: 2025-07-20]. https://oecd.ai/en/ai-principles

OpenAI. (2023a). Gpt-4 technical report [Accessed: 2025-07-19]. https://openai.com/research/gpt-4

OpenAI. (2023b). Openai text embedding models. https://platform.openai.com/docs/guides/embeddings

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. Guilford Press.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

PHINEO gAG. (2023). The iooi model: From inputs to impact [Accessed: 2025-07-20]. https://www.phineo.org/magazin/impact-orientation-and-impact-assessment

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *EMNLP 2019*. https://arxiv.org/abs/1908.10084

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson Education Limited.

SHAP. (2024). *Explainable ai with shap*. Retrieved July 21, 2024, from https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

Sun, T., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, *36*(2), 368–383. https://doi.org/10.1016/j.giq.2018.09.008

Union, E. (2024). The eu artificial intelligence act: Regulation 2024/1689. Retrieved July 3, 2025, from https://eur-lex.europa.eu/eli/reg/2024/1689

United Nations. (2024). *Sustainable development goals*. Retrieved July 21, 2025, from vahttps://sdgs.un.org/ue

UnternehmerTUM. (2023a). Guide to impact management for startups. Retrieved July 3, 2025, from https://www.unternehmertum.de/impact-management

UnternehmerTUM. (2023b). Impact guide: Wirkungsorientierung für startups. Retrieved July 3, 2025, from https://www.unternehmertum.de/impact-guide

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wirtz, B. W., Weyerer, J. C., & Mueller, C. (2020). Public administration in the age of digitalization – from e-government to smart government. *International Journal of Public Administration*, *43*(10), 832–844. https://doi.org/10.1080/01900692.2019.1619810

Appendix A

Additional Data

This appendix contains supplementary materials that support the AI-supported IMM artefact described in Chapters 4 and **??**. It includes raw project profiles, generated KPIs, audit logs, and dashboard mock-ups.

1. Example Project Profiles

Below are anonymized structured outputs from the `Profile` Pydantic model for two sample projects. These illustrate the type of information extracted from pitch decks and structured for KPI generation.


```
{
  "startup_name": "GreenFields AgTech",
  "legal_form": "GmbH",
  "founder_first_name": "Anna",
  "founder_last_name": "Schmidt",
  "founder_gender": "female",
  "startup_email": "anna@greenfields.com",
  "startup_phone": "+49 123 456 789",
  "startup_city": "Berlin",
  "startup_country": "Germany",
  "startup_postcode": "10115",
  "website": "https://greenfields.com",
  "project_beginning": "2023-03-01",
  "turnover": 250000,
  "profit": 30000,
```

```
  "employers": 5,
  "problem": "Excessive synthetic nitrogen usage in small farms",
  "vision": "Reduce fertilizer use while maintaining yields",
  "mission": "Develop sustainable precision farming tools",
  "solution": "IoT soil sensors with AI-driven recommendations",
  "social_impact": "Promote sustainable agriculture and environmental health",
  "reason": "Reduce environmental pollution and farmer costs",
  "value_1": "Sustainability",
  "value_2": "Efficiency",
  "value_3": "Innovation",
  "target_group": "Smallholder farmers in Europe"
}
```


TODO: Add 1–2 more anonymized profiles illustrating diverse sectors and project types.

—

2. Generated KPIs / Indicators

Example KPI generated from the above project profile:


```
{
  "category": "Agrar & Agrar Tech",
  "sub_category": "Sustainable Inputs",
  "goal": ["Reduce synthetic nitrogen application rate by 20 percent per hectare within 12 month
  "short_term_goal_1": "Users reduce synthetic nitrogen rate within 6 months",
  "short_term_indicator_1": "Share of users who reduced synthetic nitrogen rate compared to last
  "short_term_question_1": "Did you reduce your synthetic nitrogen application rate per hectare
  "type_of_short_term_question_1": "single_choice",
  "answer_options_short_term_question_1": ["Yes", "No", "Not applicable"],
  "measurement_method_short_term_question_1": "Self-reported comparison to baseline season recor
  "unit_method_short_term_question_1": "percent of users",
  "justification_method_short_term_question_1": "User-level rate reduction is an early signal of
  "source_method_short_term_question_1": "IRIS+ Agrochemical Use intensity; SDG 2.4",
  "long_term_goal_1": ["Users sustain a 20 percent lower synthetic nitrogen rate after 3 seasons
  "long_term_indicator_1": ["Kilograms of synthetic nitrogen applied per hectare per season"],
  "long_term_question_1": "How many kilograms of synthetic nitrogen per hectare did you apply th
```

```
  "type_of_long_term_question_1": "open_question",
  "measurement_method_long_term_question_1": "Farmer input logs normalized by field area",
  "unit_method_long_term_question_1": "kg N/ha",
  "justification_method_long_term_question_1": "Rate per area directly measures fertilizer press
  "source_method_long_term_question_1": "IRIS+ Agrochemical Use; FAO fertilizer statistics; SDG
}
```


TODO: Include 2–3 more KPIs per category to show coverage and variety. **TODO:** Add example of secondary goal handling when multiple outcomes are detected.

—

3. Human-in-the-Loop Audit Logs

> *"Short-term indicator was slightly ambiguous; refined wording to ensure farmers understand units and target."*
>
> *"SDG mapping verified: matches SDG 2.4 (Zero Hunger) and SDG 12.4 (Responsible Consumption)"*

TODO: Include anonymized full audit log table showing KPI revisions, reviewer comments, and XAI trace outputs.

—

4. Dashboard Mock-Up

Figure A.1: Prototype Impact Dashboard Showing KPI Performance and Trends

TODO: Populate dashboard with real or simulated KPI metrics. **TODO:** Add labels for public value dimensions, project comparisons, and trends.

—

5. Data Collection Instruments

Example survey question derived from KPI:

Short-term KPI: Share of users who reduced synthetic nitrogen rate **Question:** Did you reduce your synthetic nitrogen application rate per hectare this season compared to last season? **Type:** Single choice **Answer Options:** Yes / No / Not applicable **Measurement Method:** Self-reported comparison to baseline season recorded at onboarding

TODO: Add example long-term KPI survey question. **TODO:** Include one qualitative open-ended question from mission/vision evaluation.

—

6. Raw Analysis Outputs

TODO: Include anonymized sample outputs of KPI aggregation, summary statistics, or thematic analysis of qualitative data. **TODO:** Add visualizations like histograms or word clouds if available.

—

7. Ethical and Governance Documentation

- GDPR-compliant consent form template **TODO: add redacted form**.
- Notes on anonymization and secure storage **TODO: describe procedure**.
- Guidelines for human-in-the-loop oversight **TODO: include short description or table**.

—

8. Glossary and Abbreviations

- KPI – Key Performance Indicator
- IMM – Impact Measurement and Management
- SDG – Sustainable Development Goal
- LLM – Large Language Model
- XAI – Explainable Artificial Intelligence
- IRIS+ – Impact Reporting and Investment Standards

TODO: Expand glossary with pipeline-specific terms (e.g., ‘LangGraph‘, ‘PyPDF extraction‘, ‘Profile Pydantic Model‘).

—

Note: This appendix is intended to provide transparency and reproducibility for the artefact's processing and outputs, while keeping all data anonymized and compliant with privacy standards.