Design Matters: Unveiling the Impact of AI Engineering Decisions on Alignment *A Case Study in School Dropout Prediction*

Leticia Arco[0000−0002−5154−4441] and Johan Loeckx[0000−0003−4208−1411]

AI Lab, Vrije Universiteit Brussel, Pleinlaan 9, 1050 Brussels, Belgium
{larcogar,Johan.Loeckx}@vub.be
https://ai.vub.ac.be/

Abstract. AI has achieved remarkable successes, yet 80% of proof-of-concept projects fail to reach production, encounter unintended side effects, or face ethical, legal, or societal backlash. Despite the extensive research from computational and sociological perspectives, there is a significant lack of rigorous studies that connect the two fields. In this paper, we claim that a holistic and interdisciplinary understanding of alignment is needed and that rigorous AI engineering plays a crucial role to improve quality/productivity. We present a survey of the current challenges that AI engineers face while designing AI systems in practice, drawing from both academic and grey literature. We show how innovation (people, tech, data, and process), business (feasibility, desirability, and viability), epistemic (theory, formalisation, and implementation), and lifecycle perspectives intersect, interact, and impact engineering decisions. However, there are still many open research questions: what a good AI engineering methodology should look like; what is missing; how to diagnose issues given the limited understanding of problems; how to translate contextual information into actionable design decisions; how to evaluate whether design choices truly improve outcomes; how to understand interdependencies and cascading effects across stages; and how to align quality, impact, and priorities when designing a solution. We illustrate our findings in a real application of predicting school dropout.

Keywords: AI engineering · Design decisions · Holistic alignment.

1 Introduction

Artificial Intelligence (AI) has achieved remarkable successes and is receiving massive attention in research and practice. The global AI market size was valued at USD 279.22 billion in 2024 and is projected to reach USD 1,811.75 billion by 2030, growing at a compound annual growth rate (CAGR) of 35.9% from 2025 to 2030 [12]. However, 50%-80% of proof-of-concept projects fail to reach production [7], encounter unintended side effects, or face ethical, legal, or social backlash. While traditional software projects are already complex, failure-prone, and demand a broad range of expertise, AI systems add complexity through data dependencies, model uncertainty, interpretability and real-world adaptation [23].

The complexity and uniqueness of AI solutions underscore the urgent need for dedicated research in AI engineering, a discipline focused on designing, developing, deploying, and maintaining AI systems that are not only functional but also robust, explainable, contestable, ethically aligned, and scalable in real-world contexts. Advancing AI engineering practices is critical not only to improving the reliability and accountability of AI systems, but also to achieve alignment technology, societal values, ethical principles, and stakeholder needs, ensuring that their deployment meets both technical and societal expectations. This is a multifaceted concept, where the epistemic framing of AI engineering decisions, along with the insight that alignment extends beyond technical correctness to include interpretability, usability, and social acceptability, is crucial.

Most definitions of AI engineering describe it as a multidisciplinary field focused on building scalable, operational AI systems by integrating software, data, and systems engineering with human-centered design, emphasizing both the technical infrastructure and the design of intelligent systems that function effectively in real-world enterprise contexts [10, 22]. Taking these foundations further, we conceptualize AI engineering as a socio-technical practice that encompasses not only technologies and methods but also the people, processes, and data involved in the design, development, deployment, and use of AI systems. AI engineering comprises a continuous and collaborative process, where decisions span multiple disciplines, epistemic abstraction levels, and lifecycle stages requiring careful alignment between technical components, ethical values, contextual and domain knowledge, and organizational dynamics. Here, alignment extends beyond traditional AI safety to serve as a guiding design principle: ensuring that decisions consistently reflect stakeholder needs, ethical expectations, intended societal outcomes, and previous decisions made.

Despite increasing interest and extensive research in the field, AI engineering continues to struggle with implementing responsible practices in real-world system design. A central problem addressed in this paper is the lack of rigorous interdisciplinary studies that account for the complexity of responsible design decisions throughout the AI engineering process. Current approaches often treat design decisions as purely technical, overlooking the broader contextual and socio-ethical implications. This narrow framing limits the ability to develop systems that are truly aligned with societal values and stakeholder needs. To address this gap, this paper proposes an interdisciplinary holistic paradigm for responsible AI engineering, integrating four key perspectives—business, innovation, epistemic, and lifecycle—and positioning them at the core of technical design rather than as peripheral considerations. By adopting this paradigm, we aim to guide AI engineers toward more informed, context-sensitive, and ethically grounded decisions throughout the system lifecycle.

The key highlights of this paper are: (1) emphasize the relevance of transparency, alignment, and accountability throughout the AI design process, (2) raise awareness and attract attention on the impact of design decisions, (3) provide a rationale for deeper research on responsible AI engineering, (4) illustrate the effect of design choices through concrete examples, and (5) expose the need

for more transparent design decisions. While primarily aimed at AI engineers, the paper is also relevant for decision-makers who rely on AI outputs; as understanding the consequences of design choices is essential for responsible use.

The following is a critical reflection on AI engineering, highlighting its unique characteristics, identifying key research gaps and limitations in current design practices. Next, we present the holistic paradigm for responsible AI engineering and illustrate its relevance through a case study on school dropout prediction. The paper concludes with open research questions and final reflections.

2 Critical Reflections on AI Engineering

The development of AI systems introduces fundamental challenges that go beyond those addressed by traditional software engineering. AI engineering addresses the complex intersection between AI and software development, exposing limitations in traditional engineering practices when applied to AI systems. Unlike conventional software where the change occurs primarily in code, AI systems evolve along three axes: code, data, and models, each introducing its own risks and needs [24]. Furthermore, AI systems must address cross-cutting concerns such as ethics, legal compliance, user trust, contestability, and security throughout their entire lifecycle, not just during model development [17]. AI projects include activities such as model training and performance drift monitoring, none of which align well with traditional engineering tools, timelines, or architectural patterns [24]. The susceptibility of AI to specialized adversarial attacks requires additional custom-made defensive strategies [24]. Moreover, AI engineering involves multidisciplinary collaboration among experts, complicating integration and communication [23]. Ultimately, the challenge lies not just in creating effective models, but in establishing robust processes that support responsible AI deployment throughout the entire lifecycle of the system [7].

2.1 AI Engineering: Computational and Sociological Perspectives

As described above, AI systems possess unique characteristics that challenge the assumptions and practices of traditional software engineering. These differences demand new engineering approaches to ensure responsible design, development, and deployment of AI technologies [17]. In response, a growing body of research has emerged around responsible AI engineering, with some efforts focusing on computational and technical aspects [15, 4, 9, 6, 5, 23, 27, 16, 24], while others address broader sociological and ethical dimensions [3, 11, 21, 25].

Recently, efforts have been made to bridge sociological and computational perspectives in AI research. De et al. [8] advocate an integrated interdisciplinary approach to better align ethical principles with practical software engineering processes. Himmelreich and Köhler [14] argue that responsibility gaps in AI should be addressed through conceptual engineering, focusing on redefining responsibility based on its societal function to resolve underlying conceptual disagreements. Lu et al. [19, 18, 20] argue that responsible AI has focused too

narrowly on algorithms, neglecting system-level concerns such as vague ethical requirements, limited risk assessment, and weak post-deployment monitoring, and propose addressing these gaps through lifecycle-integrated ethics, design patterns, architectural guidance, and a Responsible AI Pattern Catalog to support multilevel governance, trustworthy processes, and responsible-AI-by-design throughout the system lifecycle. Sanderson et al. [26] examine the implementation of high-level AI principles through four key areas —requirements, design, implementation, and deployment—, adapted from software engineering practices, situating them within an ethical framework that includes governance, competency development, and stakeholder communication. Harbi et al. [13] propose a comprehensive framework to integrate responsible design patterns into Machine Learning (ML) pipelines to mitigate risks and ensure ethical development, deployment, and post-deployment of AI systems. Stahl [28] criticizes AI ethics for overlooking socio-technical complexity and proposes the concept of meta-responsibility to promote system-level ethical accountability and highlights conditions for fostering responsible AI ecosystems.

2.2 Key Gaps in Responsible AI Engineering

Despite extensive research, AI engineering still faces challenges in implementing responsible design practices effectively. Based on the literature review, the analysis of real-world use cases, insights from over 35 applied projects, and consulting experience, we identified key gaps organized into four categories (see Fig. 1).

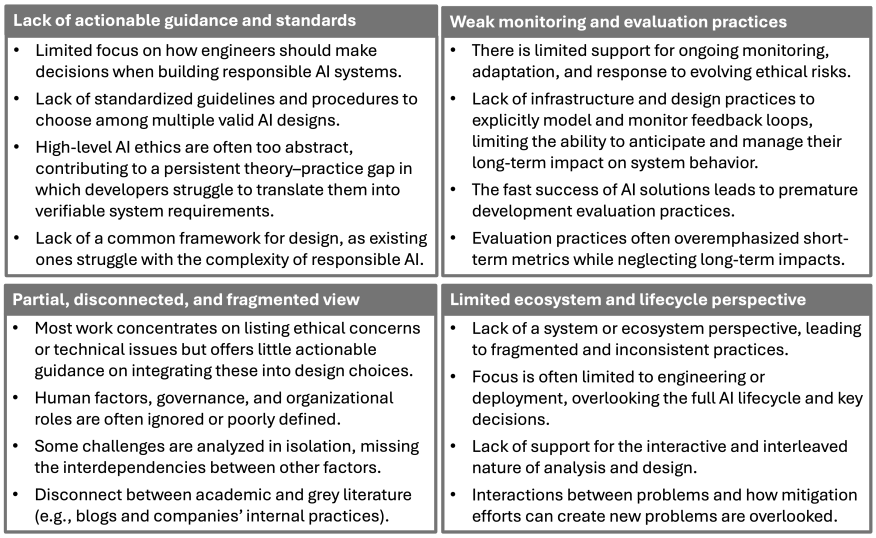

Fig. 1. Key gaps in current responsible AI engineering include the absence of actionable guidance and standards, partial, fragmented and disconnected approaches, a narrow focus lacking ecosystem and lifecycle perspective, and insufficient support for monitoring and evaluation.

Responsible AI engineering involves much more than addressing ethical concerns or technical issues; it requires navigating a complex landscape of interdependent factors that influence design decisions. Although ethics plays a critical role, making the "right" engineering choices also requires alignment between AI foundations, technology capabilities, domain-specific knowledge, contextual constraints, and stakeholder needs. For instance, ensuring that an AI system is trustworthy and useful means not only following fairness or transparency guidelines but also understanding how technical options interact with real-world practices, data limitations, methods requirements, and the expectations of diverse users. A responsible approach must therefore integrate these dimensions to guide decisions that are both technically sound and socially meaningful.

2.3 Limitations in Current AI Design Practices

A key issue in current responsible AI engineering, discussed in this paper, is the lack of introspection and rigorous studies examining the complexity of responsible design decision-making throughout the engineering process. Even seemingly technical choices (e.g., handling unbalanced classes, adjusting the loss function, or applying upsampling) require careful consideration beyond mere algorithmic optimization. Drawing on our experience in applied AI projects and exchange with AI engineers, and in line with the gaps identified in AI engineering, we highlight four main pitfalls in AI design practices, as shown in Fig. 2.

We argue that informed and responsible design choices require a holistic, interdisciplinary understanding of alignment that integrates technical, contextual, and socio-ethical considerations. This approach enhances the rigor of AI engineering while promoting higher system quality and productivity.

3 A Holistic Understanding of Alignment

This section introduces an integrative paradigm aimed at advancing responsible AI system design. While prior discussions highlighted critical gaps in AI engineering and insufficient attention to the broader implications of design decisions, we now turn to how these challenges and limitations can be addressed.

Our proposal is built upon empirical engagement, theoretical analysis, and practical experience. It draws from interactions with over 200 companies, alongside a review of established innovation and business strategy models (e.g., Business Model Canvas), which emphasize value creation and organizational alignment, as well as established standards and frameworks (e.g., ISO), where lifecycle thinking is widely recognized. These foundations led to the integration of innovation, business, and lifecycle perspectives. However, we identified a critical gap: the absence of a perspective addressing the theoretical grounding, formal reasoning, representational integrity, and aligned implementation of AI systems. To address this, we introduced the epistemic perspective as an original contribution. Together, these four dimensions constitute a minimal yet sufficient set for engineering AI systems responsibly.

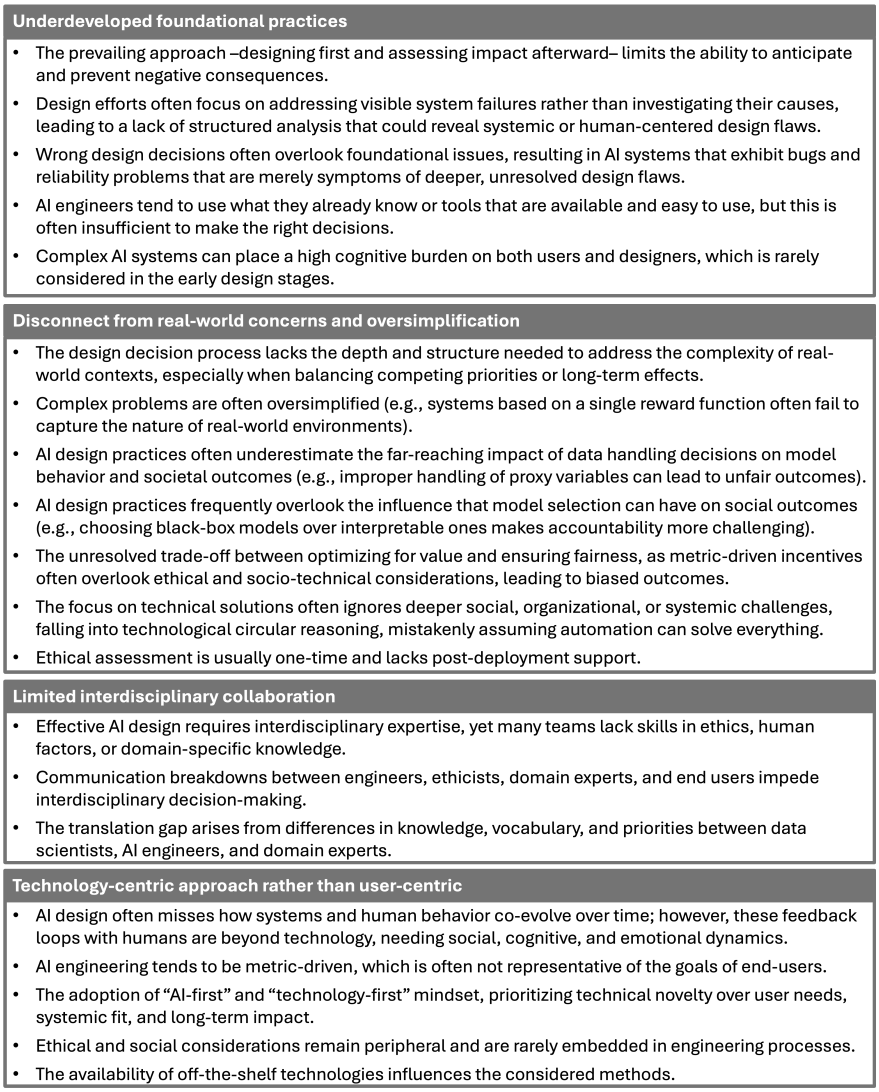

Fig. 2. Limitations in AI design practices include weak foundational methods, limited interdisciplinary collaboration, and a disconnect from real-world concerns, compounded by a technology-centric approach that often overlooks user needs and societal context.

Fig. 3 presents the new holistic interdisciplinary paradigm that integrates four essential perspectives —business, innovation, epistemic, and lifecycle— as core components of responsible AI engineering. Rather than treating these dimensions as external or secondary to technical design, this paradigm positions them as active forces shaping engineering decisions throughout all stages of the AI development lifecycle. Their intersection, interaction, and influence on design

and implementation choices are illustrated through examples from a case study on predicting school dropout. These examples illustrate key points, but do not cover all possible scenarios that could emerge from the case study.

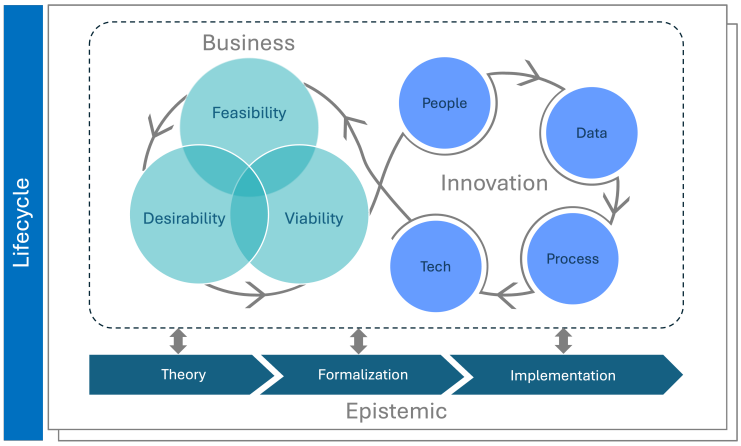

Fig. 3. A new paradigm for responsible AI engineering integrating business, innovation, lifecycle, and epistemic perspectives to guide design decisions.

The school dropout prediction use case is ideal to illustrate the relevance of the proposed paradigm because it involves complex social impacts that demand ethical and fairness considerations. It engages diverse stakeholders (e.g., teachers, policymakers, educational advisory members, students, parents) with varying priorities, requiring alignment of all perspectives. The use case presents significant data challenges, highlighting the importance of epistemic factors in modeling decisions. Its evolving nature showcases lifecycle concerns such as model monitoring and updating. Furthermore, it involves critical trade-offs between accuracy and fairness, emphasizing the need to balance technical and ethical factors. Finally, it offers opportunities for innovation through early interventions and personalized support, demonstrating how integrated perspectives influence AI design throughout the development lifecycle.

While the school dropout use case illustrates responsible AI engineering challenges for traditional machine learning systems, a Large Language Model (LLM)-based use case might introduce a distinct set of challenges. In LLM applications, alignment becomes an even bigger problem because computers do not "speak" a language; human instructions must first be translated, creating an initial semantic gap, while a second gap emerges during the system's generation of responses, accompanied by additional security concerns. In LLMs, a single channel (the prompt) serves for programming and input, whereas in conventional systems these roles are separate, and users are not allowed to directly program system functionality, which is a key difference. More broadly, AI has introduced new

programming paradigms —from rule-based programming, functionality defined throughout labeling, to reinforcement learning shaped by feedback state of the world, and now to LLMs controlled via natural language— which introduces new AI engineering challenges, leaving open research questions.

The work being done to make AI systems more robust, responsible, contestable, explainable, accountable, ethical, fair, and human-centered contributes significantly to improving alignment. We must learn from these fields to enhance our work. To incorporate these advances into our paradigm, desirable principles and concepts from Explainable AI, Contestable AI, Human-Centered AI, and Trustworthy AI can be translated into technical specifications, design constraints, and system requirements at the definition level, guiding implementation across the AI lifecycle in our paradigm. A major challenge lies in ensuring that these principles remain aligned with their formal implementation; in this regard, related practices and operationalization proposals can be highly relevant. For example, contestable AI features [1] —such as interactive control over automated decisions and explanations of system behavior— can be translated into specifications or requirements. Contestable AI practices are therefore valuable for operationalizing these obligations within engineering decisions across the AI lifecycle. Similarly, in addressing legal aspects within the proposed paradigm, our approach does not take a stance to which regulation or policies but wants to provide a framework for people to understand and verify whether their compliant (broader than legal: security, medical, privacy, ethics, etc.). Legal and policy obligations are translated into specifications or requirements, raising the central question of how such obligations materialize in design choices.

3.1 The Business Perspective

AI engineering decisions are not made in a vacuum; they must align with business realities and goals. This underscores the fundamental role of the business perspective in responsible AI engineering and its inclusion in our paradigm. The emphasis is placed on how feasibility, viability, and desirability shape responsible AI development. These dimensions define the practical, economic, and human-centered boundaries within which choices must be made to ensure systems are not only functional but also sustainable, valuable, and ethically grounded.

Fig. 4 highlights how business-oriented dimensions collectively support and guide responsible AI practices. Feasibility focuses on what can be done with available resources directing design choices toward practical and implementable solutions, while viability ensures that engineering decisions are economically sustainable, generating long-term business value. Desirability highlights the need to design systems that users find appealing, intuitive, valuable, and ethically acceptable, focusing on user experience, accessibility, and ethical alignment.

In the context of school dropout prediction, several examples highlight the importance of the feasibility dimension in shaping design decisions. For instance, suppose that the IT system cannot support real-time data streaming, engineers may decide against implementing a real-time dropout prediction model and instead opt for batch processing. Another example is when algorithms fail to ac-

curately predict dropout risk due to insufficient data, often caused by high data collection costs, which raises viability considerations. It is also very likely that the data do not provide a full picture of the problem, due to ethical constraints or decisions made by policymakers, both of which relate to desirability considerations. This example illustrates not only the impact of feasibility but also the interplay among various business dimensions.

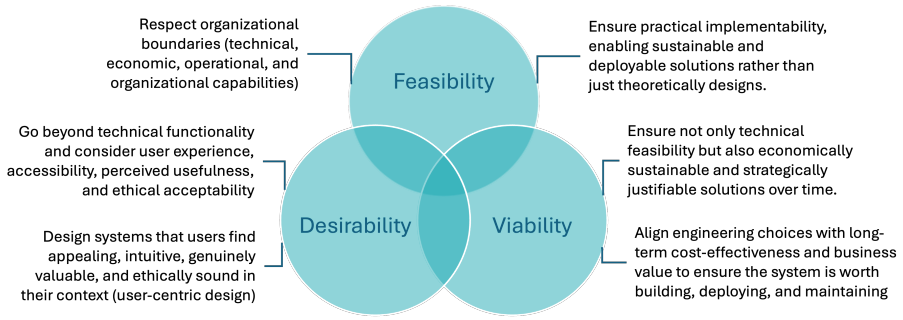

Fig. 4. Business-oriented dimensions must guide AI engineering decisions and interact to shape responsible AI design: feasible systems must also be viable and desirable to achieve successful, sustainable, and widely adopted AI solutions.

While a highly complex model using diverse data sources could improve prediction accuracy, the costs of data collection (e.g., it needs to happen by a third party), storage, and ongoing analysis would be very high, illustrating the relevance of viability in this application domain. If advisors struggle to interpret the outputs or find them irrelevant to their workflows, the system may be ignored. From a desirability dimension, engineering teams should focus on user-centered design, use interpretable methods, and build explanations that meet user needs.

The EU AI act forbids classifying emotional state and classifies education-related AI as high-risk, while GDPR imposes strict data-protection and transparency requirements. Within our approach, such legal mandates can be translated into different design considerations: in terms of desirability, certain features may be prohibited altogether; in terms of viability, implementing adequate governance and oversight may be too costly; and in terms of feasibility, avoiding specific features to meet legal constraints could significantly impact model performance. By integrating these considerations into early scoping, data handling, model design, and deployment, our approach ensures that compliance is systematically embedded rather than post-hoc checks.

3.2 The Innovation Perspective

This section introduces four key dimensions that shape AI engineering decisions from the innovation perspective: people, process, technology, and data. Each

dimension plays a distinct role in influencing how AI systems are designed, developed, deployed, and maintained, highlighting the interplay between human, organizational, technical, and informational factors throughout the system lifecycle, as shown in Fig. 5. The following dropout prediction examples illustrate the relevance of these dimensions for practicing responsible AI engineering.

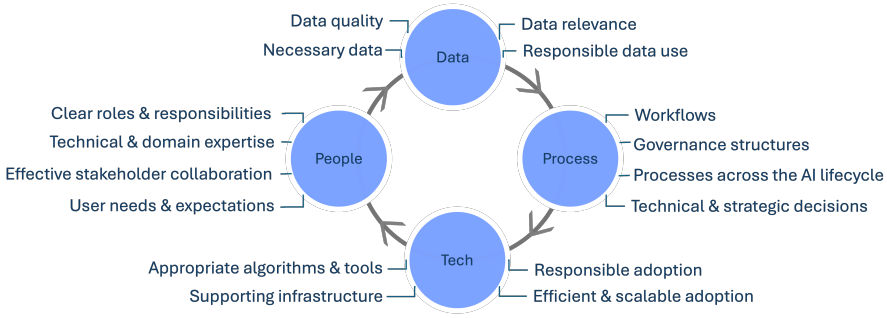

Fig. 5. Overview of the Innovation perspective in AI engineering, highlighting how the core dimensions shape and inform design decisions. The People dimension emphasizes the roles, expertise, collaboration, and user needs of individuals and teams involved in AI development, use, and management. The Process dimension highlights workflows and governance structures that shape the AI lifecycle and support technical and strategic decision-making. The Technology dimension informs the selection of algorithms, methods, tools, and infrastructure, as well as the capabilities required for efficient, scalable, and responsible adoption. The Data dimension defines how data supports system objectives and ensures quality, relevance, and responsible use.

People. Designing a system for dropout risk prediction must account for the diverse needs and perspectives of multiple stakeholders (e.g., students, teachers, and policymakers). Each group may require different types of information, levels of detail, and modes of interaction. For instance, students may need personalized understandable feedback; teachers may require classroom-level insights to support interventions; and policymakers could look for aggregated trends to inform strategic decisions. These differences influence choices around user interfaces, data access, design decisions, explanation mechanisms, and privacy settings in the system's architecture. The level of technical proficiency among stakeholders can determine how to present outputs; requiring interpretable models. There are also challenges related to the knowledge, practices, preferences, expertise, psychology, and cognitive tendencies of AI engineers and developers (e.g., they may default to familiar techniques when designing solutions). Human factors influence design decisions throughout the entire AI system lifecycle and must be acknowledged in responsible AI engineering.

Process. In the context of dropout prediction, understanding who generated the data and under what circumstances is critical. For instance, a grade of

7 might reflect different standards or expectations depending on the teacher, which introduces variability that impact model reliability. Similarly, the timing of data collection plays a significant role, data gathered before COVID-19 may not reflect the current learning environment or student behavior. If predictions are expected post-COVID, engineers must carefully evaluate the relevance and validity of historical data, possibly adjusting features or retraining models to account for shifts in context, learning modalities, and student needs.

Technology. Designing a dropout prediction system requires technologies that ensure secure integration with existing student information systems while complying with data protection regulations. Privacy-preserving techniques directly shape architectural choices, model selection, and deployment tools, as some approaches may not be compatible with privacy constraints or could compromise data protection if not properly integrated. For instance, privacy rules may block centralizing student data, requiring federated learning where schools train locally and share only model updates. This preserves privacy but adds challenges (e.g., data heterogeneity, model convergence, and secure communication) that impact architecture, algorithms, and infrastructure choices. In addition, if the school's IT infrastructure is limited (e.g., no access to cloud services or limited computing power), the system must be designed to run efficiently on local machines with minimal processing requirements.

Data. If national education policy prohibits the use of socioeconomic data (e.g., family income, parental education) in predictive analytics to prevent discrimination, AI engineers must exclude these features when developing the model. This policy constraint directly impacts feature selection and limits the types of data that can be collected or used. As a result, engineers may need to identify alternative proxy variables, which introduces additional design challenges because such proxies could inadvertently introduce new biases or reduce model accuracy. These trade-offs influence subsequent decisions across the ML pipeline, including model training, validation, interpretation, and deployment.

3.3 The Epistemic Perspective

Building on the concept presented by Almeida et al. [2], AI solutions are not just technical artifacts but institutional architectures that shape and organize complex systems of interaction between humans and machines. They embody epistemic choices about what knowledge is relevant, how it should be structured, and how decisions are made. They are not just decision tools, but also dynamic socio-technical constructs that encode values, mediate power, structure our understanding of knowledge and action in society, establish formal and informal rules that guide behavior, shape interactions, and distribute power.

The epistemic perspective is crucial for responsible AI engineering because it not only encompasses the nature, sources, limitations, and validity of knowledge within AI systems, but also addresses how these systems represent, reason about, and communicate that knowledge, directly influencing accountability,

transparency, and trustworthiness in AI design, development, and deployment. Epistemic aspects can be understood and addressed through three interconnected layers: theory, formalization, and implementation. Fig. 6 describes key elements in these layers that critically influence AI design choices and highlights how epistemic considerations shape responsible AI engineering.

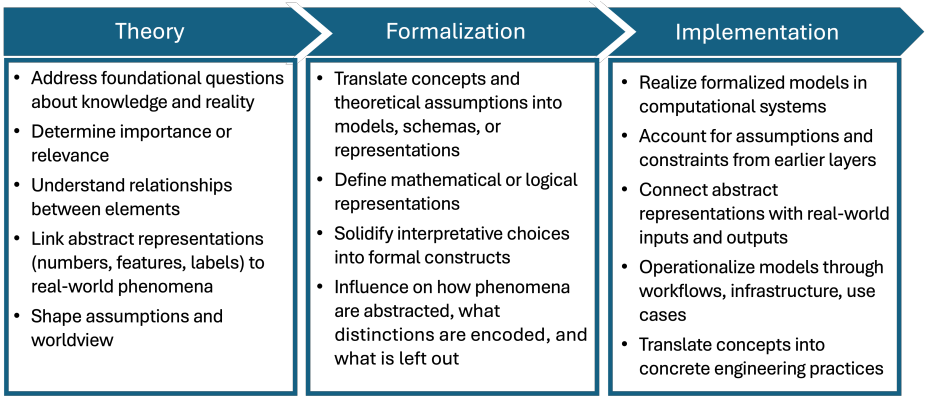

Fig. 6. Overview of the Epistemic perspective in AI system design, illustrating the three interconnected layers. The Theory layer addresses foundational questions about knowledge, relevance, relationships between elements, and the mapping of abstract representations to real-world phenomena, shaping the assumptions and worldview underlying the AI system. The Formalization layer translates these theoretical concepts into operational models, schemas, and representations, defining mathematical or logical encodings, solidifying interpretative choices, and embedding epistemic values. The Implementation layer focuses on the practical realization of these formalized models within computational systems, connecting abstract representations to real-world inputs and outputs, operationalizing epistemic commitments through workflows, infrastructure, and use cases, influencing system behavior and user impact.

Theory. At the theoretical level, dropout is defined as a student permanently leaving an educational program before completion. However, different stakeholders or institutions may have varying definitions based on enrollment status, academic performance, or engagement levels. Clarifying this concept is essential to guide system goals. But understanding what dropout means is only part of the challenge; it is equally important to conceptualize how dropout risk is expressed, whether as a probability, a categorical risk level, or a composite score. This defines the nature of knowledge the AI system aims to represent and influences subsequent modeling approaches. Furthermore, it is also crucial to know who has the authority and expertise to define and interpret dropout risk; for instance, which actors (e.g., academic advisors or external educational bodies) should be primarily responsible for the decision. Understanding these roles helps define how knowledge is structured and how decision boundaries are established.

In general, the meaning of features directly influences decisions at multiple stages of the ML pipeline, including feature selection, model design, and evaluation. For example, a grade of 7 may hold different significance depending on the teacher or vary for students with different learning needs (e.g., a student with dyslexia compared to one without). Properly accounting for the semantics of features can affect processes like normalization, but neglecting these nuances can lead to misleading conclusions or biased predictions. Even more challenging could be defining teacher quality, which can influence school dropout rates. Even more so is determining how to represent this complex concept. Its interpretation often varies depending on personal perspectives and its interaction with other factors. Furthermore, there is no universally accepted metric for measuring teacher quality, and it can fluctuate over time. These challenges have a significant impact on system design and the resulting outcomes. Therefore, it is essential to contextualize features using both technical methods and domain expertise to ensure accurate and fair modeling.

Formalisation. The way we formalize the problem has a significant epistemic impact on system design. If dropout is defined as a binary outcome (e.g., dropout vs. enroll), the task is formalized as a binary classification problem, while if the risk is defined as high, low, or medium, it could be formalized as an ordinal problem problem. In contrast, if risk is defined as a continuous score indicating the likelihood or degree of risk, it becomes a regression problem. This choice affects the entire modeling pipeline, from how the target variable is constructed to which algorithms are suitable, how uncertainty is handled, and how results are communicated to stakeholders. For instance, a regression-based score might be more useful for prioritizing interventions, while classification might be better for triggering binary decisions (e.g., flag or no flag). Each approach implies different assumptions about what counts as knowledge, what level of precision is meaningful, and how actionable the output is for end users such as teachers or counselors.

Implementation. Defining dropout as a binary classification problem directly shapes key design decisions in the system (e.g., choosing appropriate classifiers). This influences what features are prioritized and how outputs are integrated into intervention workflows, such as generating alerts for school counselors. Ultimately, the decision to model dropout in this way drives the entire architecture of the predictive system, from data preparation to model deployment and user interaction. In addition, choice of programming language will influence performance and thus how/where it can be deployed (e.g., C(++) could be deployed in some custom-made lightweight device while Python will need a "normal" computer and will be much slower).

3.4 The Lifecycle Perspective

The lifecycle perspective is central to our paradigm, recognizing that AI engineering goes beyond model training or algorithm selection and demands a comprehensive, integrated approach to designing, developing, deploying, and main-

taining systems—encompassing every stage where decisions influence behavior, quality, and societal impact. By adopting a lifecycle view, AI engineers are better positioned to identify and address risks and challenges across different phases, even more important because early-stage AI design decisions have impacts on later phases (e.g., mitigation strategies applied at one stage may introduce unintended biases or trade-offs in others), underscoring the critical importance of understanding and managing interactions across the entire development lifecycle. In addition, continuous feedback loops and iterative refinement become essential for adapting to real-world changes and improving system robustness over time. In summary, this perspective underscores the need for system-level alignment across all phases to ensure responsible, sustainable, and trustworthy AI.

In the context of dropout prediction, early decisions about which student data to collect (e.g., attendance, grades, gender, or nationality) directly affect the quality, accuracy, and fairness of the dropout prediction model, influencing choices in subsequent stages. For example, selecting attendance as a key feature can introduce bias if the definition or recording of absences varies across institutions. More broadly, models must remain effective and fair across diverse and changing contexts: a model trained on centralized data from one region may underperform in others, and reliance on outdated data, such as pre-COVID records, can degrade performance over time. These challenges highlight the importance of feedback loops and lifecycle-aware monitoring to ensure AI systems remain reliable and unbiased throughout their deployment.

4 Open Research Questions

Despite growing awareness of the importance of responsible AI engineering, several critical questions remain unanswered. In particular, there is a pressing need for clearer systematization to guide AI engineers in making informed and context-sensitive design decisions. This section outlines open research questions aimed at defining what a robust AI engineering methodology should include and effectively integrate alignment, transparency, and interdisciplinary considerations throughout the development lifecycle.

- How would a good AI engineering methodology look like?
- What is currently missing or underdeveloped in AI engineering practices?
- How to diagnose problems if we even do not have an overview or good understanding of the problems and we lack good or complete definitions?
- How design choices depend on context, use, and domain knowledge?
- How can we reliably detect issues early in the design process, before they propagate across pipeline stages?
- How can AI engineers make informed decisions when designing solutions given the huge number of alternatives?
- What methods can be used to translate contextual and domain-specific information into actionable design decisions in AI workflows?
- How can AI engineers really assess whether a design decision leads to meaningful performance improvements and aligns with expected outcomes?

- How can AI engineers and data scientists be effectively guided to evaluate their solutions from an alignment lens? What kind of systematization is needed for this guidance?
- What are the interdependencies and cascading effects of decisions made in early stages to later stages in a ML pipeline?
- How do mitigation strategies implemented in some phases potentially introduce new challenges in subsequent phases?
- What frameworks or best practices can help anticipate and manage the impact design decisions across the full lifecycle of an AI solution?
- How to align quality, impact, priorities, and pragmatism when designing a solution to be good enough?
- How can semantic misalignment between human intent and LLM interpretation be formally modeled and mitigated across diverse use contexts?
- What engineering principles are needed to responsibly manage the dual role of prompts as both input and program in LLM-based systems?

The research questions highlight the urgent need for clearer systematization that can guide practitioners in navigating trade-offs, assessing alignment, and integrating domain knowledge throughout the lifecycle of AI systems. We believe this is not a challenge that should be left to individual sectors or practitioners to solve independently. Rather, the AI community—spanning technical and non-technical disciplines alike—must actively engage in developing shared methods, tools, and principles to advance responsible AI engineering.

5 Conclusions

This paper has argued that responsible AI engineering requires a shift from fragmented, technically centered practices to more holistic and interdisciplinary approaches. Despite increasing interest in the field, many AI engineering practices still fall short in accounting for the full complexity of responsible design decisions. In particular, decisions are often made in isolation from the broader business, social, ethical, and epistemic contexts in which AI systems operate. This results in systems that may be technically sound but misaligned with societal values, stakeholder expectations, or long-term impacts.

The proposed paradigm for responsible AI engineering foregrounds four essential and interconnected perspectives: business, innovation, epistemic, and lifecycle. Rather than treating these as external constraints, we position them as core elements of the design process. Through a case study on school dropout prediction, we illustrated how these perspectives can interact and influence engineering decisions across different stages of system development.

We hope this work contributes to ongoing efforts in the academic and industry communities to reframe AI engineering as a socio-technical practice. Our aim is not to offer a definitive blueprint, but to open up space for reflection, dialogue, and further research on how to better support AI engineers and decision-makers in making transparent, aligned, and accountable design choices, serving as a starting point for broader interdisciplinary discussion and inviting collaboration to address the socio-technical challenges of AI design.

Acknowledgments. The authors thank the VUB AI Applied Research Team for their constructive suggestions and the anonymous reviewers for their valuable feedback, which helped improve the quality and clarity of this manuscript.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alfrink, K., Keller, I., Kortuem, G., Doorn, N.: Contestable AI by design: Towards a framework. Minds and Machines p. 613–639 (2023). https://doi.org/10.1007/s11023-022-09611-z
2. Almeida, V., Mendonça, R.F., Filgueiras, F.: Thinking of algorithms as institutions. Communications of the ACM **68**, 20–23 (1 2025). https://doi.org/10.1145/3680411
3. Bano, M., Zowghi, D., Shea, P., Ibarra, G.: Investigating responsible AI for scientific research: An empirical study. Responsible AI for Science (2023)
4. Belani, H., Vukovic, M., Car, Z.: Requirements engineering challenges in building AI-based complex systems. In: Proceedings of the International Requirements Engineering Conference Workshops (REW 2019). pp. 252–255. IEEE Computer Society (9 2019). https://doi.org/10.1109/REW.2019.00051
5. Biswas, S., Wardat, M., Rajan, H.: The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large. In: Proceedings of the International Conference on Software Engineering. vol. 2022-May, pp. 2091–2103. IEEE Computer Society (2022). https://doi.org/10.1145/3510003.3510057
6. Bosch, J., Crnkovic, I., Olsson, H.H.: Engineering AI systems: A research agenda (6 2020), https://doi.org/10.48550/arXiv.2001.07522
7. Bosch, J., Olsson, H.H., Brinne, B., Crnkovic, I.: AI engineering: Realizing the potential of AI. IEEE Software **39**, 23–27 (2022). https://doi.org/10.1109/MS.2022.3199621
8. De, M., Leça, M., Bento, F., De, R., Santos, S.: Responsible AI in the software industry: A practitioner-centered perspective https://arxiv.org/pdf/2412.07620
9. Fischer, L., Ehrlinger, L., Geist, V., Ramler, R., Sobiezky, F., Zellinger, W., Brunner, D., Kumar, M., Moser, B.: AI system engineering: Key challenges and lessons learned (2020), https://doi.org/10.3390/make3010004
10. Gartner: Information technology glossary: AI engineering. https://www.gartner.com/en/information-technology/glossary/ai-engineering (2025), accessed: 2025-07-15
11. Goellner, S., Tropmann-Frick, M., Brumen, B.: Responsible artificial intelligence: A structured literature review (3 2024), http://arxiv.org/abs/2403.06910
12. GrandViewResearch: Artificial intelligence market (2025 - 2033). https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market (2025), accessed: 2025-07-15
13. Harbi, S.H.A., Tidjon, L.N., Khomh, F.: Responsible design patterns for machine learning pipelines (6 2023), http://arxiv.org/abs/2306.01788
14. Himmelreich, J., Köhler, S.: Responsible AI through conceptual engineering. Philosophy and Technology **35** (9 2022). https://doi.org/10.1007/s13347-022-00542-2
15. Leslie, D.: Understanding artificial intelligence ethics and safety (6 2019), http://arxiv.org/abs/1906.05684

16. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B.: Trustworthy AI: From principles to practices. ACM Computing Surveys **55** (9 2023). https://doi.org/10.1145/3555803
17. Lu, Q., Zhu, L., Whittle, J., Michael, J.B.: Software engineering for responsible AI. Computer **56**, 13–16 (4 2023). https://doi.org/10.1109/MC.2023.3242055
18. Lu, Q., Zhu, L., Xu, X., Whittle, J.: Responsible-AI-by-design: a pattern collection for designing responsible AI systems (9 2022), http://arxiv.org/abs/2203.00905
19. Lu, Q., Zhu, L., Xu, X., Whittle, J., Douglas, D., Sanderson, C.: Software engineering for responsible AI: An empirical study and operationalised patterns (11 2021), http://arxiv.org/abs/2111.09478
20. Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., Jacquet, A.: Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering. ACM Computing Surveys **56** (7 2024). https://doi.org/10.1145/3626234
21. Löhr, G.: If conceptual engineering is a new method in the ethics of AI, what method is it exactly? AI and Ethics **4**, 575–585 (5 2024). https://doi.org/10.1007/s43681-023-00295-4
22. MIT: MIT professional education: What is artificial intelligence engineering? https://professionalprograms.mit.edu/blog/technology/artificial-intelligence-engineering/ (2025), accessed: 2025-07-15
23. Nahar, N., Zhou, S., Lewis, G., Kästner, C.: Collaboration challenges in building ML-enabled systems: Communication, documentation, engineering, and process (2 2022), http://arxiv.org/abs/2110.10234
24. Paleyes, A., Urma, R.G., Lawrence, N.D.: Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys **55** (7 2023). https://doi.org/10.1145/3533378
25. Papagiannidis, E., Mikalef, P., Conboy, K.: Responsible artificial intelligence governance: A review and research framework (6 2025). https://doi.org/10.1016/j.jsis.2024.101885
26. Sanderson, C., Lu, Q., Douglas, D., Xu, X., Zhu, L., Whittle, J.: Towards implementing responsible AI. In: Proceedings of the International Conference on Big Data (Big Data 2022). pp. 5076–5081. Institute of Electrical and Electronics Engineers Inc. (2022). https://doi.org/10.1109/BigData55660.2022.10021121
27. Soklaski, R., Goodwin, J., Brown, O., Yee, M., Matterer, J.: Tools and practices for responsible AI engineering (1 2022), http://arxiv.org/abs/2201.05647
28. Stahl, B.C.: Embedding responsibility in intelligent systems: from AI ethics to responsible AI ecosystems. Scientific Reports **13** (12 2023). https://doi.org/10.1038/s41598-023-34622-w