Ethical Principles Application in the Development of AI Systems: a Systematic Mapping Study

Helena Cristo Martins[1], **Cristiane Aparecida Lana**[1,2], **Maria Lúcia Bento Villela**[1]

[1]Departamento de Informática – Universidade Federal de Viçosa (UFV)
Viçosa – MG – Brasil

[2]Universidade Federal de Lavras (UFLA-Paraíso)
São Sebastião do Paraíso – MG – Brasil

{helena.martins, maria.villela}@ufv.br, cristiane.lana@ufla.br

Abstract. *The widespread adoption of Artificial Intelligence (AI) in high-stakes domains demands the effective operationalization of ethical principles. This study identifies and classifies solutions that support the integration of ethics across the AI system lifecycle. A Systematic Mapping Study was conducted, analyzing 102 studies (2019–2024) and categorizing artifacts by ethical principles, approach types, and development stages. The results show a predominance of conceptual and procedural artifacts in early lifecycle phases, limited support for later stages, an emphasis on Justice and Explicability, and comparatively low attention to Autonomy, revealing persistent gaps between principles and practice.*

1. Introduction

The widespread adoption of Artificial Intelligence (AI) across domains has intensified ethical concerns related to transparency, fairness, privacy, accountability, and autonomy [Chancellor 2023]. As AI systems increasingly shape high-stakes decisions, aligning their development with human values has become both a societal and technical imperative [Morley et al. 2020, Shneiderman 2020].

Despite the proliferation of ethical frameworks proposed by governments, industry, and international organizations, a persistent gap remains between abstract principles and their practical integration into AI development processes [Morley et al. 2020]. Developers and designers often lack structured support to translate high-level guidance into actionable decisions across the AI lifecycle.

Previous secondary studies have addressed this challenge [Morley et al. 2020, Prem 2023, de Paula Porto et al. 2025, Ortega-Bolaños et al. 2024, Batool et al. 2025]. However, the AI landscape has evolved substantially, marked by the rapid expansion of generative AI, regulatory consolidation such as the [Act 2024], and advances in Human-Centered AI (HCAI) research [Ozmen Garibay et al. 2023, Capel e Brereton 2023]. These developments have stimulated new lifecycle-oriented methods, governance mechanisms, and monitoring approaches aimed at operationalizing ethical principles.

These developments have introduced a new generation of artifacts aimed at operationalizing AI ethics; however, it remains unclear how they are distributed across the AI lifecycle, which ethical principles they address, and where critical gaps persist. Existing secondary studies predate or only partially capture these shifts, leaving the current landscape undercharacterized.

This study aims to characterize the artifacts currently available to support ethical integration throughout the AI lifecycle, identifying the principles they address and the stages they cover. To this end, a Systematic Mapping Study (SMS) was conducted, adopting a lifecycle-based analytical perspective. It: (i) characterizes artifacts by type and approach; (ii) analyzes the ethical principles and issues they address; and (iii) maps their applicability across stages of algorithmic development. By integrating insights from prior reviews through a unified lifecycle-oriented lens, this work helps bridge the gap between ethical principles and practical AI development. The study aligns with GranDIHC-BR 2025–2035, particularly GC2 (Ethics and Responsibility) [Rodrigues et al. 2024] and GC6 (Implications of Artificial Intelligence in HCI) [Duarte et al. 2024, Pereira et al. 2024].

The remainder of this paper is organized as follows: Section 2 presents the analytical foundations of the study; Section 3 presents related work; Section 4 details the research methodology; Section 5 presents the results; the key findings are discussed in Section 6; and finally, Section 7 presents the conclusions and future work.

2. Analytical Foundations: Ethical Principles, Approaches and AI Lifecycle

Human-Centered Artificial Intelligence (HCAI) seeks to align AI development with human values and responsible innovation, integrating ethical considerations throughout the system lifecycle rather than treating them as post-hoc evaluations [Shneiderman 2022, Xu 2019]. Over the past decade, principle-based ethical frameworks have proliferated, with studies highlighting convergence around recurring dimensions [Jobin et al. 2019] and examining their interpretation in practice [Khan et al. 2023, de Paula Porto et al. 2025].

Given the absence of a universally standardized list of principles, this study adopts the five principles proposed by [Floridi et al. 2018] as a consolidated analytical framework: *beneficence*, *non-maleficence*, *autonomy*, *justice*, and *explicability*. *Beneficence* concerns promoting societal well-being; *non-maleficence*, preventing harm; *autonomy*, preserving meaningful human control and agency; *justice*, ensuring fairness and non-discrimination; and *explicability*, guaranteeing transparency and accountability. This framework synthesizes recurrent ethical dimensions identified in prior research [Floridi et al. 2018, Jobin et al. 2019, Khan et al. 2023, de Paula Porto et al. 2025], offers conceptual clarity, and supports comparability in lifecycle-oriented analysis.

To structure the analysis, two complementary lenses are adopted. First, following [Prem 2023], ethical operationalization mechanisms are categorized into seven approach types: *Summaries*; *Notions* (e.g., frameworks, checklists, datasheets, metrics); *Procedures* (process models and standards); *Code* (algorithmic methods and software tools); *Infrastructure* (datasets and communities); *Education* (training materials); and *Ex-post assessments and agreements* (audits and regulatory mechanisms). These categories capture instruments ranging from conceptual to technical-operational levels.

Second, the lifecycle perspective builds on the seven-stage Machine Learning pipeline from [Morley et al. 2020]: *business and use-case development* (problem definition and AI suitability assessment); *design* (translation of objectives into technical requirements); *training and test data procurement* (dataset collection and documentation); *building* (model implementation); *testing* (validation and failure identification); *deploy-*

ment (real-world release); and *monitoring* (continuous tracking of performance and impacts). As ethical challenges vary across these stages, they serve as the reference model for mapping artifact applicability.

By integrating (i) Floridi's ethical principles [Floridi et al. 2018], (ii) Prem's categories of approaches [Prem 2023], and (iii) Morley's lifecycle segmentation [Morley et al. 2020], this study establishes a multidimensional analytical framework. This integration enables systematic classification of artifacts by principle, intervention type, and lifecycle stage, supporting a structured analysis of how ethical guidance is translated into practical mechanisms across AI development.

3. Related Work

Several studies have examined the ethical challenges of Artificial Intelligence, addressing principles, governance frameworks, practical tools, and strategies for responsible development. Secondary studies, in particular, have sought to clarify how high-level ethical principles are translated into operational mechanisms within AI system development.

A foundational contribution is provided by [Morley et al. 2020], who proposed a lifecycle-oriented typology grounded in the five principles of beneficence, non-maleficence, autonomy, justice, and explicability. Their work mapped publicly available tools and methods to stages of the machine learning pipeline, revealing an uneven distribution across principles and development phases. By explicitly articulating ethical principles with lifecycle stages, they advanced the shift from defining "what" ethical AI entails to examining "how" it can be implemented.

[Prem 2023] expanded this perspective by systematically analyzing the artifacts identified by [Morley et al. 2020], i.e., more than 100 ethical AI frameworks, process models, and proposed remedies for implementing AI ethics in practice. The study categorizes approaches by governance mechanisms and intervention types, offering a broad typological overview of how ethical concerns are addressed.

Focusing on requirements engineering, [de Paula Porto et al. 2025] conducted a systematic review of techniques and tools for eliciting and managing ethical requirements in AI systems. Although this work provides valuable insights into early lifecycle practices, it does not extend the analysis across the full development pipeline. Empirical research has also contributed to the debate. For example, [Khan et al. 2023] examined practitioners' and lawmakers' perspectives, identifying recurring dimensions such as fairness, transparency, accountability, privacy, and safety, while highlighting differences between stakeholder groups. However, such studies do not provide structured mappings of artifacts across lifecycle stages.

More recently, large-scale secondary reviews have expanded the landscape. The systematic review by [Ortega-Bolaños et al. 2024] cataloged hundreds of tools for developing and assessing AI systems, organizing them by lifecycle stages and ethical dimensions, and identifying persistent imbalances in coverage. While comprehensive, its primary focus is tool inventory and classification rather than analyzing how different types of approaches operationalize ethical principles. Similarly, [Batool et al. 2025] conducted a meta-analysis of AI governance mechanisms, emphasizing actor- and stage-specific responsible AI tools and the need for clearer alignment between ethical objectives, development phases, and organizational roles.

Despite these advances, gaps remain. Prior studies have either emphasized lifecycle, principle alignment [Morley et al. 2020], typological classification [Prem 2023], domain-specific requirements [de Paula Porto et al. 2025], or large-scale cataloging and governance mapping [Ortega-Bolaños et al. 2024, Batool et al. 2025]. Given the rapid evolution of generative AI, regulatory consolidation, and the proliferation of new lifecycle-oriented artifacts, there is a need for an updated and methodologically transparent mapping that integrates ethical principles, types of operational mechanisms, and development stages.

This study addresses this need through a Systematic Mapping Study (SMS) that organizes findings according to stages of the AI development lifecycle. The mapping examines how artifacts address ethical principles (*beneficence*, non-*maleficence*, *autonomy*, *justice*, and *explicability*), their approach categories, and their applicability across lifecycle stages, complementing and extending prior secondary studies with a structured and contemporary perspective.

4. Research Method

This study adopts a Systematic Mapping Study (SMS) to identify, classify, and analyze artifacts aimed at operationalizing ethical principles throughout the AI system development lifecycle. The mapping was conducted following the SMS guidelines proposed by [Petersen et al. 2015], which informed the overall design, execution, and reporting of the mapping process. Complementarily, general recommendations for systematic studies from [Kitchenham e Charters 2007] were adopted to guide the definition of the review protocol[1], ensuring transparency, rigor, and replicability.

4.1. Planning the Systematic Mapping Study

The systematic mapping covered studies published between January 2019 and May 2024. This time frame was established as the cut-off period in the review protocol to maintain methodological consistency and avoid iterative expansion of the dataset during analysis. As emphasized by [Kitchenham e Charters 2007] and [Petersen et al. 2015], systematic studies necessarily rely on clearly defined temporal boundaries to preserve internal validity and replicability. Although AI ethics research continues to evolve rapidly, the selected period captures a recent and representative snapshot of lifecycle-oriented artifacts emerging in response to contemporary technological and regulatory developments.

4.1.1. Research Questions

As the research goal is to map solutions to address ethics in the development of AI systems, the main research question (RQ), aligned with the aforementioned goal, is: ***What artifacts are available to support developers and designers in reflecting on and applying ethical principles throughout the AI development lifecycle?***

To address this RQ, we defined three specific questions (SQ) to better characterize the solutions, shown in Table 1.

[1]See "1-Systematic Mapping Study Protocol" at Zenodo repository `https://doi.org/10.5281/zenodo.19477437`

Table 1. Specific Questions

Specific Question	Examples
SQ1 What types of approaches do the solutions consist of?	*Notions, Procedures, Code, Infrastructure, Education, Ex-post assessment and agreement*
SQ2 Which ethical principles do these solutions address?	*Beneficence, Non-Maleficence, Autonomy, Justice, Explicability*
SQ3 At which stage(s) of the AI system lifecycle can they be applied?	*Business and use-case development, System design, Training and test data procurement, Building, Testing, Deployment, Monitoring*

4.1.2. Search Strategy

To ensure methodological transparency, the search string was constructed following the PICOC framework (Population, Intervention, Comparison, Outcomes, Context), as recommended by [Kitchenham e Charters 2007].

The PICOC elements were operationalized as follows: Population (P) - AI systems and their development processes; Intervention (I) - Ethical artifacts, tools, frameworks, methods, and practices; Comparison (C) - Not applicable, as the study aims at mapping and classification rather than comparative evaluation [Petersen et al. 2015]; Outcomes (O): Operationalization of ethical principles in development activities; Context (C): Artificial Intelligence and the AI system lifecycle.

The string construction followed a multi-step validation process: (i) Initial derivation based on prior secondary study [Morley et al. 2020]; (ii) Expert consultation with HCI researchers; (iii) Pilot searches to assess coverage and precision; (iv) refining the terms by identifying relevant synonyms; and (v) Refinement to balance recall and specificity. The final search string was defined as:

("software development" OR "algorithm development") AND ethic* AND ("artificial intelligence" OR "machine learning")

While the inclusion of development-related terms may appear closer to Software Engineering discourse, this choice reflects the study's lifecycle-oriented perspective and its focus on artifacts that support practical ethical integration in AI system projects, rather than purely conceptual or normative discussions.

The search was conducted in the following digital libraries: **ACM Digital Library**[2], **IEEE Xplore Digital Library**[3], **Scopus**[4], and **Springer**[5]. The combination of these databases was intended to minimize disciplinary bias and ensure coverage of both technical and design-oriented research communities. In particular, ACM Digital Library was included to capture contributions emerging from the HCI community, which may frame ethical concerns in terms of design, interaction, or user-centered approaches rather than strictly software engineering terminology. Additionally, we included **arXiv** [6] to capture recent gray literature, given the rapidly evolving nature of AI ethics research.

[2] https://dl.acm.org/
[3] https://ieeexplore.ieee.org
[4] http://scopus.com
[5] https://link.springer.com/
[6] https://arxiv.org/

Digital libraries were selected to ensure broad coverage of AI, HCI, and Software Engineering research, prioritizing multidisciplinary databases with wide indexing and international scope, in line with established guidelines [Kitchenham e Charters 2007, Petersen et al. 2015]. The Brazilian Computer Society Open Library (SOL/SBC) was not queried separately, as its proceedings are partially indexed in databases already included (e.g., Scopus and ACM DL).

4.1.3. Studies Selection

Specific selection criteria were applied across three screening rounds. Studies were included if they presented an approach that developers or designers could practically apply to integrate ethical considerations into AI system development.Studies were excluded if they met at least one of the following criteria: [EC1] Not aligned with the study objective; [EC2] Not a scientific article or book chapter (e.g., tutorial, editorial, poster); [EC3] Duplicate or superseded version of the same research; [EC4] Secondary study compiling other works; [EC5] Not written in English or Portuguese; [EC6] Full text unavailable.

As an exploratory systematic mapping study focused on classification and coverage [Petersen et al. 2015], no formal quality assessment was conducted.

4.2. Conducting the Systematic Mapping Study

To execute the process of paper selection, the search string defined above was used in the selected digital libraries on May 09, 2024, and the set of retrieved publications was stored in the Parsif.al[7] for further analysis. The selection process involved the participation of three researchers to reduce the bias caused by a single researcher applying the research method. To ensure the reliability of the results, each paper was analyzed by one researcher under the supervision of two other researchers, and disagreements were discussed until a consensus was reached.

4.2.1. Screening Procedure

The studies were filtered in three stages, with selection criteria applied at each step:

- **Preliminary reading (1st filter):** Titles, abstracts, and keywords were screened to remove clearly unrelated studies; doubtful cases were retained for further review.
- **Diagonal reading (2nd filter):** Introductions, main sections, and conclusions were examined to assess relevance to the research questions; uncertain cases progressed to the next stage.
- **Complete reading (3rd filter):** Full texts were reviewed, and the final set of studies was selected for data extraction.

Following the defined protocol, 3442 studies were retrieved, and 3208 remained after duplicate removal. After the three screening stages (preliminary, diagonal, and full-text reading), 102 primary studies were selected for data extraction. Figure 1 summarizes the selection process. The details of the studies analyzed are available with open access[8].

[7] `https://parsif.al/`

[8] Available at `https://doi.org/10.5281/zenodo.19477437` ("2-Table of Studies").

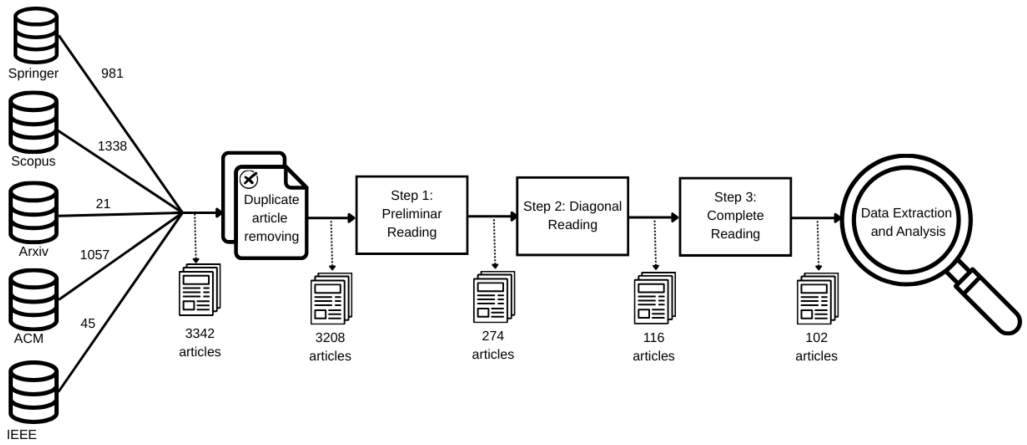


Figure 1. Selection process

4.2.2. Data Extraction, Classification and Synthesis

Data extraction was conducted in MS Excel using a structured sheet[9] designed to capture demographic data and the information required to answer the secondary research questions. The identified solutions were classified through three complementary analytical frameworks: Prem's categories of ethical AI approaches [Prem 2023], Floridi's five ethical principles [Floridi et al. 2018], and Morley's lifecycle stages [Morley et al. 2020]. This combined classification enabled a multidimensional mapping of solutions across types of approaches, ethical principles, and development phases. As this study constitutes an exploratory systematic mapping focused on classification and coverage [Petersen et al. 2015], no formal quality assessment of the included studies was performed.

4.3. Threats to Validity

To enhance the reliability of our Systematic Mapping Study (SMS) and minimize potential biases introduced by the authors [Kitchenham e Charters 2007] [Wohlin et al. 2012], we identified key threats to validity and implemented specific mitigation measures:

- **Search bias:** Use of multiple databases and structured Boolean search strings.
- **Selection bias:** Predefined inclusion and exclusion criteria and documented multi-stage screening.
- **Classification bias:** Use of a consolidated ethical framework [Floridi et al. 2018] and clearly defined lifecycle stages to standardize artifact categorization [Morley et al. 2020].
- **Researcher bias:** Protocol definition prior to execution and consensus-based resolution of ambiguous cases.
- **Temporal limitation:** The rapid evolution of AI research may lead to the publication of new artifacts after the defined cut-off date (May 2024). While the temporal boundary ensures methodological rigor and reproducibility, it may not capture subsequent developments. Future research may extend this mapping to incorporate

[9]Detailed information about the Excel sheet can be found at `https://doi.org/10.5281/zenodo.19477437`("3-Data Extraction Form").

studies published after the analyzed period and reassess potential shifts in artifact distribution across principles and lifecycle stages.

4.4. Ethical Considerations

This study did not involve human participants and therefore did not require ethics committee approval. The systematic literature mapping followed the ethical guidelines established by the Brazilian Computer Society (SBC).

5. Results

The next sections present the results of our SMS. Section 5.1 shows an overview of the main studies used to answer our RQ, which are analyzed in detail in Sections 5.2, 5.3, and 5.4. It is important to explain that one study can be included in more than one category, so the total number of categories may be greater than the total number of studies (n = 102).

5.1. Characterization of Primary Studies

Between January 2019 and May 2024, we identified 102 publications on this topic. Figure 2 shows steady growth in research, mainly due to advances in AI, its use in important sectors, and concerns about ethics, such as transparency and algorithmic bias. In 2021, the number of studies increased significantly, along with discussions about regulations such as the European Union Artificial Intelligence Act [Act 2024] and the UNESCO ethical principles [UNESCO 2021]. In 2023, research grew again, especially due to debates about generative AI and its social impacts.

Most studies originated from the United States (28), followed by Finland (12), Germany and Australia (8 each), and the Netherlands (7), with other countries contributing fewer than six studies each[10]. The predominance of the USA and Europe reflects strong R&D investment and academia–industry collaboration, while countries such as Australia and Brazil demonstrate growing engagement in AI ethics research.

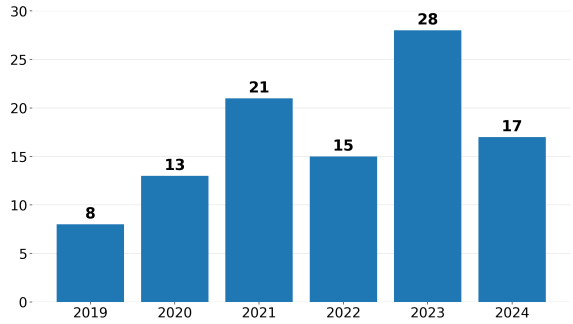


Figure 2. Studies per year

5.2. SQ1: What types of approaches do the solutions consist of?

The 102 primary studies were classified using Prem's taxonomy [Prem 2023], which includes seven categories: *Summaries, Notions, Procedures, Code, Infrastructure, Education*, and *Ex-post assessments and agreements* (see Section 2). The *Summaries* category was excluded because it includes only general descriptions and not practical solutions.

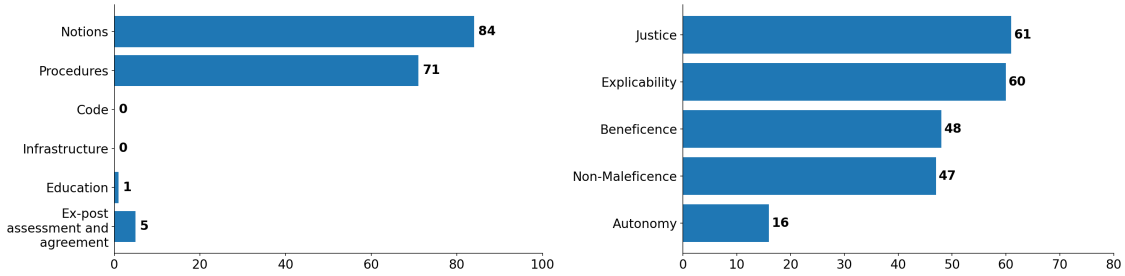
[10]Detailed information on the studies by country can be found at `https://doi.org/10.5281/zenodo.19477437`("4-Studies by Country").

Figure 3a shows that *Notions* (n = 84) and *Procedures* (n = 71) were the most common categories, together representing 96% of the 161 occurrences[11]. Many solutions (n = 56) were classified in both categories, which means they combine conceptual ideas with practical guidance. For example, S97 discusses how to adapt an AI fairness checklist representing *Notions*. In *Procedures*, S1 extends the ECCOLA method with a structured deployment model to support ethical AI practices. Combining both, S13 presents a responsible AI framework that connects ethical principles with practical steps across the AI lifecycle.

In contrast, only five solutions were classified as *Ex-post assessment and agreement*, showing little focus on monitoring AI after development. For example, S18 proposes a three-layer auditing framework for large language models to evaluate ethical risks after deployment. Only one study was classified as *Education*, and none were found in the *Code* or *Infrastructure* categories. In the *Education* category, S82 includes discussions about algorithmic bias in an introductory computer science course to help students understand fairness and social impacts in algorithm design.

5.3. SQ2: Which ethical principles do these solutions address?

The analysis of ethical principles shows an uneven distribution among the five dimensions. *Justice* (n = 61) and *Explicability* (n = 60) are the most frequent (60% of studies), reflecting a strong emphasis on fairness and transparency. *Beneficence* (n = 48) and *Non-Maleficence* (n = 47) are also common, while *Autonomy* is the least addressed (n = 16), suggesting limited support for user agency. Figure 3b presents these results. Classification followed the conceptual definitions of [Floridi et al. 2018]. In particular, Beneficence was assigned to solutions promoting positive societal outcomes, and Non-Maleficence to those preventing harm or mitigating risks; solutions addressing both were classified as multi-principle contributions. A detailed mapping, including the translation of ethical principles into tangible system requirements, is provided in the supplementary material to support traceability and reproducibility[12].”



(a) Frequency of Approach Categories **(b) Frequency of Ethical Principles**

Figure 3. Distribution of Approaches and Ethical Principles in solutions

For example, in a case study on candidate screening, S72 investigates methods for building and auditing fair algorithms by identifying and mitigating sources of disparate impact across demographic groups, illustrating how procedural and metric-based

[11]Detailed information on the studies by category can be found at `https://doi.org/10.5281/zenodo.19477437` (“5-Studies by Category”).

[12]See 6-Studies by Ethical Principle” and 10-Translating Ethical Principles into Tangible System Requirements” at Zenodo: `https://doi.org/10.5281/zenodo.19477437`

techniques can be applied to uphold the principle of *Justice* in real-world AI systems. Similarly, S11 presents a framework implementation for accountable and explainable artificial intelligence that links model internals with user-facing explanations, demonstrating how system design and documentation can enhance the transparency and interpretability of complex models and support *Explicability* in practical AI deployments.

S7 proposes a reflective decision-making approach embedding ethical impact analysis into professional practice, emphasizing developers' responsibility to anticipate societal consequences and minimize unintended harm (*Beneficence* and *Non-Maleficence*). S90 describes a pragmatic regulatory framework emphasizing voluntary compliance and human oversight to preserve independent decision-making (*Autonomy*), while also addressing *Justice* through equitable regulatory assurance across critical sectors and *Explicability* through transparent, auditable criteria for AI regulation and stakeholder communication.

Overall, the findings indicate that ethical principles are predominantly addressed in combination rather than in isolation. Most solutions integrate multiple ethical dimensions, with *Explicability* and *Justice* emerging as the most frequent co-occurring pair. The joint consideration of *Beneficence* and *Non-Maleficence*, often in association with other principles, is also recurrent. This pattern suggests that the analyzed AI solutions adopt a multidimensional ethical perspective, seeking simultaneously to mitigate potential harms and to promote fairness, transparency, and user-centered benefits[13].

5.4. SQ3: Stage(s) of the AI system lifecycle where solutions can be applied

The distribution of ethical tools across the AI lifecycle shows a clear concentration in early phases, especially *Design* and *Training and test data procurement*, with less attention to the other stages[14]. This pattern indicates that ethical integration is prioritized in initial decisions but less sustained during implementation and real-world operation.

As shown in Figure 4, the *Design* phase has the highest concentration of approaches (n = 99), highlighting the focus on embedding ethical values when translating requirements into technical specifications. As an example, S74 demonstrates how design checklists support alignment between user needs and ethical principles.

The *Training and test data procurement* stage is the second most addressed (n = 86). This aligns with widespread recognition of the pivotal role of data quality and representativeness in mitigating bias and ensuring fairness. Tools in this phase often focus on ethical compliance in dataset usage, including bias detection, dataset documentation, and fairness metrics, all aimed at preventing discrimination and safeguarding sensitive information. A representative study here is S72, which proposes data auditing techniques to detect and remediate harmful patterns before model training.

In contrast, phases such as *Building* (n = 20), *Testing* (n = 17), and *Deployment* (n = 12) are significantly less represented. For example, S4 illustrates an ethical construction approach that integrates unit-level fairness tests into model pipelines, highlighting initial efforts in this area. Similarly, S73 proposes an end-to-end internal algorithmic auditing

[13]Detailed graphic showing the frequency of ethical principle combinations is available at `https://doi.org/10.5281/zenodo.19477437` ("7-Ethical Principles addressed by combinations").

[14]Detailed information on the studies by stage of algorithmic development can be found at `https://doi.org/10.5281/zenodo.19477437` ("8-Studies by Stage of algorithmic development").

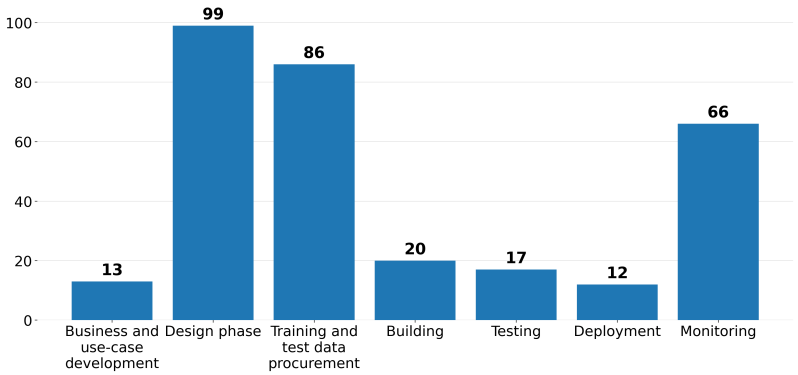


Figure 4. Distribution of Stages of Algorithm Development

framework designed to systematically evaluate machine learning systems for ethical risks before deployment, and S22 proposes a set of explainability-oriented artifacts for software systems, aiming to support developers and evaluators in assessing and communicating explainability aspects during system evaluation and deployment, thereby helping to ensure transparency and comprehensibility when systems are put into operation. Although these contributions demonstrate important steps toward ethical practice, their relatively low frequency suggests a need for more extensive support at these lifecycle points.

The *Monitoring* stage (n = 66) demonstrates a growing interest in ongoing evaluation of AI systems' ethical performance after deployment. Tools in this category include runtime fairness monitors, drift detectors, and logging frameworks that support accountability and long-term assessment. For instance, S52 proposes a method that encourages continuous ethical reflection and documentation of decisions, enabling ongoing oversight of ethical considerations as systems evolve in practice.

Taken together, these findings indicate that ethical AI support is currently skewed toward the initial design and data phases, with comparatively less emphasis on sustained ethical reflection and adaptation during *Building*, *Testing*, and *Deployment*, which increases again in the *Monitoring* stage.

6. Discussion

By answering the research question (RQ) *"What artifacts are available to support developers and designers in reflecting on and applying ethical principles throughout the AI development lifecycle?"*, this study identifies both advances and limitations in the development of AI ethics. The findings indicate a predominance of solutions centered on *Justice* and *Explicability*, with moderate attention to *Beneficence* and *Non-Maleficence*, while *Autonomy* remains comparatively underexplored. This imbalance reveals a field largely driven by concerns with fairness and transparency, yet still limited in addressing the broader spectrum of ethical principles in a systematic manner.

The analysis further shows that most contributions are concentrated in the categories of *Notions* and *Procedures*, particularly within early lifecycle phases such as *Design* and *Training and Test Data Procurement*, as shown in Figure 5. Although these approaches provide essential conceptual and process-oriented guidance, they often lack integration with development environments and runtime infrastructures. In contrast, more

technical and practice-oriented categories, including *Code*, *Infrastructure*, *Education*, and *Ex-post Assessment and Agreement*, remain underrepresented. This distribution suggests limited tool support, scarce automated assessment mechanisms, insufficient pedagogical resources, and fragile accountability structures. Moreover, ethical guidance tends to diminish in mid and late-stage phases, *Building*, *Testing*, *Deployment*, and *Monitoring*, thereby hindering the sustained and systematic integration of ethical principles throughout the AI system lifecycle.

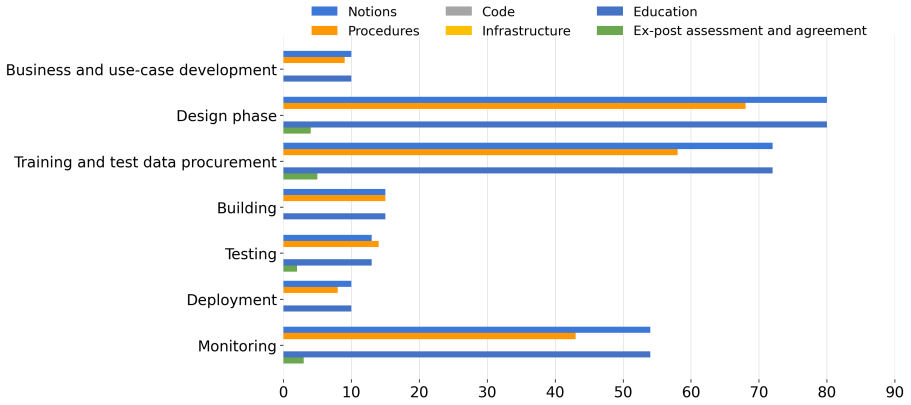


Figure 5. Cross-distribution of ethical approaches across AI lifecycle stages

These findings align with prior research while introducing relevant conceptual refinements. Consistent with [Morley et al. 2020], our results confirm the persistent gap between high-level ethical principles and their practical operationalization across AI development stages. Similarly, [Prem 2023] and subsequent systematic reviews [Batool et al. 2025, Ortega-Bolaños et al. 2024] emphasize that most frameworks privilege concerns such as explicability, fairness, and accountability, remaining largely conceptual and concentrated in early lifecycle phases. Our analysis corroborates this pattern, highlighting the continued predominance of abstract guidance over integrated, tool-supported solutions.

Nevertheless, the typology[15] proposed in this study advances the field by offering a more comprehensive distribution of ethical principles across lifecycle stages. Unlike prior frameworks that associate specific principles with discrete phases, a systematic cross-analysis of the mapped artifacts revealed that principles such as *Autonomy* and *Beneficence* recur across multiple stages — from design decisions to monitoring mechanisms — reflecting their foundational role throughout the system lifecycle. Furthermore, *Explicability* and *Justice* emerged as transversal concerns distributed across nearly all stages, suggesting that fairness and transparency must be continuously maintained rather than treated as discrete checkpoints. As illustrated in Figure 6, this distribution challenges stage-specific frameworks and supports a more integrated ethical perspective grounded in an extensive and updated literature analysis.

Several factors may explain the observed gaps. Embedding ethical evaluation into automated workflows requires specialized tooling and clear metrics, which are currently scarce. Without standardized tests to evaluate ethical issues such as fairness and privacy,

[15] See “9-Tipology” at Zenodo repository `https://doi.org/10.5281/zenodo.19477437`.

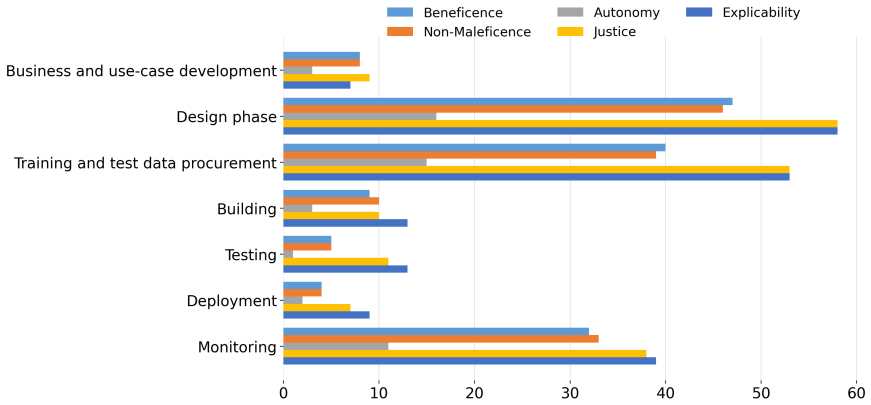


Figure 6. Distribution of ethical principles across lifecycle stages

developers may struggle to incorporate ethical checks or detect ethical degradation after deployment. Moreover, the scarcity of educational artifacts suggests that many practitioners lack training or confidence to apply ethical guidance, further impeding adoption.

From a practical perspective, these gaps matter. Without robust Testing and Monitoring support, AI systems may exhibit ethical issues in deployment that go undetected, undermining reliability and stakeholder trust. The lack of integration with standard development environments also discourages adoption, as practitioners are less likely to adopt tools that require significant additional effort. Finally, limited attention to Autonomy suggests that users' ability to exercise meaningful control remains insufficiently supported.

The findings highlight key research directions: advancing executable ethical mechanisms, integrating ethics into later lifecycle stages, establishing standardized evaluation frameworks, and strengthening autonomy-oriented design. Despite progress toward operationalization, ethics remains insufficiently embedded in AI systems, requiring better alignment between normative frameworks, technical infrastructures, and lifecycle-wide validation practices.

7. Conclusion

This study presents a Systematic Mapping Study examining how ethical principles are operationalized in AI system development. Based on the analysis of 102 primary studies (2019-2024), the mapping identified structural trends and persistent gaps in translating high-level ethical principles into concrete development practices.

The results indicate a strong concentration of ethical support in early lifecycle stages, particularly *Design* and *Training and Test Data Procurement*, while *Building*, *Testing*, and *Deployment* remain comparatively under-supported. Although *Monitoring* has gained attention, many solutions emphasize governance-oriented mechanisms rather than embedded technical artifacts. In parallel, *Notions* and *Procedures* dominate the landscape, whereas *Code* and *Infrastructure* level contributions are scarce. Regarding ethical principles, *Justice* and *Explicability* prevail, while *Autonomy* remains underexplored.

This study advances prior work by integrating ethical principles, approach categories, and lifecycle stages into a unified framework, revealing uneven distributions across mechanisms and phases. It also contributes to the GranDIHC-BR 2025–2035 agenda

(GC2 and GC6) by providing a structured assessment of ethical integration in AI engineering practices.

From an HCI perspective, the predominance of *Notions* and *Procedures* leaves designers with conceptual guidance rather than actionable tools, while the underrepresentation of Autonomy-oriented artifacts reflects insufficient support for user control and participatory engagement. The scarcity of *Testing* and *Monitoring* artifacts further limits post-deployment ethical evaluation. These gaps call for HCI-informed contributions — participatory design methods, user-centered evaluation frameworks, and accessibility-aware mechanisms — to operationalize ethical principles across the lifecycle.

Future work includes the development of a structured catalog of tools and mechanisms aligned with the proposed typology, as well as an extension of the SMS to incorporate publications from 2025, enabling an updated assessment of emerging trends and potential shifts in the distribution of ethical artifacts across lifecycle stages and principles.

Overall, although the field has progressed toward operationalization, ethics is not yet a fully integrated engineering property. Advancing this requires executable mechanisms, lifecycle-wide integration, and stronger validation to ensure sustained ethical performance in practice.

8. Acknowledgements

The ChatGPT was used to assist in the writing and revision process, with the purpose of improving clarity and linguistic accuracy.

References

Act, E. A. I. (2024). The eu artificial intelligence act.

Batool, A., Zowghi, D., e Bano, M. (2025). Ai governance: a systematic literature review. *AI and Ethics*, pages 1–15.

Capel, T. e Brereton, M. (2023). What is human-centered about human-centered ai? a map of the research landscape. In *Proceedings of the CHI 2023*, pages 1–23.

Chancellor, S. (2023). Toward practices for human-centered machine learning. *Communications of the ACM*, 66(3):78–85.

de Paula Porto, D., Prado, R. D. C. V., dos Santos Marques, G., Serrano, A. L. M., de Mendonça, F. L., e Canedo, E. D. (2025). Ethical requirements in the age of artificial intelligence: A systematic literature review. *(SBSI 2025)*, pages 663–672.

Duarte, E. F., T. Palomino, P., Pontual Falcão, T., Lis Porto, G., e Portela, Carlos e Francisco Ribeiro, D. e. N. A. e. A. Y. e. S. M. e. G. A. e. M. T. A. (2024). GranDIHC-BR 2025-2035 - GC6: Implications of Artificial Intelligence in HCI: A Discussion on Paradigms, Ethics, and Diversity, Equity and Inclusion. In *Proceedings of the IHC 2024*, New York, NY, USA. Association for Computing Machinery.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28:689–707.

Jobin, A., Ienca, M., e Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399.

Khan, A. A., Akbar, M. A., Fahmideh, M., Liang, P., Waseem, M., Ahmad, A., Niazi, M., e Abrahamsson, P. (2023). Ai ethics: an empirical study on the views of practitioners and lawmakers. *IEEE Transactions on Computational Social Systems*, 10(6):2971–2984.

Kitchenham, B. e Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering.

Morley, J., Floridi, L., Kinsey, L., e Elhalal, A. (2020). From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. *Science and engineering ethics*, 26(4):2141–2168.

Ortega-Bolaños, R., Bernal-Salcedo, J., Germán Ortiz, M., Galeano Sarmiento, J., Ruz, G. A., e Tabares-Soto, R. (2024). Applying the ethics of ai: a systematic review of tools for developing and assessing ai-based systems. *Artificial Intelligence Review*, 57(5):110.

Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S. M., Garibay, I., Grieman, K., et al. (2023). Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, 39(3):391–437.

Pereira, R., Darin, T., e Silveira, M. S. (2024). GranDIHC-BR: Grand Research Challenges in Human-Computer Interaction in Brazil for 2025-2035. In *Proceedings of the IHC 2024*, New York, NY, USA. Association for Computing Machinery.

Petersen, K., Vakkalanka, S., e Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18.

Prem, E. (2023). From ethical ai frameworks to tools: a review of approaches. *AI and Ethics*, 3(3):699–716.

Rodrigues, K. R. d. H., Carvalho, L. P., Pimentel, M. d. G. C., e Freire, A. P. (2024). GranDIHC-BR 2025-2035 - GC2: Ethics and Responsibility: Principles, Regulations, and Societal Implications of Human Participation in HCI Research. In *Proceedings of the IHC 2024*, New York, NY, USA. ACM.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31.

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

UNESCO, C. (2021). Recommendation on the ethics of artificial intelligence.

Wohlin, C., Runeson, P., Host, M., Ohlsson, M. C., Regnell, B. j., e Wessln, A. (2012). *Experimentation in software engineering*. Springer Publishing Company, Incorporated.

Xu, W. (2019). Toward human-centered ai: a perspective from human-computer interaction. *Interactions*, 26(4):42–46.