Benchmarking Education on the Ethical Aspects of Artificial Intelligence: Integrating Empathy into AI Ethics Training

Enrico Barbierato[1,*,†], Alice Gatti[1,†] and Marco Gribaudo[2,†]

[1]Università Cattolica del Sacro Cuore, Dipatimento di Matematica e Fisica, via Garzetta 48, 25123 Brescia, Italy
[1]Università Cattolica del Sacro Cuore, Dipatimento di Matematica e Fisica, via Garzetta 48, 25123 Brescia, Italy
[2]Dip. di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, via Ponzio 34/5, 20133 Milano, Italy

Abstract

Artificial Intelligence (AI) systems play a crucial role in decision-making processes across critical domains, raising urgent concerns about fairness, accountability, transparency, and societal impact. Education on the ethical aspects of AI is therefore essential for preparing developers, policymakers, and citizens to navigate these challenges. Yet existing initiatives vary widely in scope and depth, and there is no established framework for evaluating their effectiveness. This paper proposes a structured benchmark for AI ethics education, defined by measurable criteria that encompass comprehensive content coverage, diverse pedagogical strategies, practical skill development, and—distinctively—empathy cultivation, grounded in neuroscientific findings on mirror neurons. The benchmark is further illustrated through fallibility scenarios that can undermine ethical competence, such as superficial treatment of ethics, cultural blind spots, and the empathy gap, each paired with corrective actions within an iterative improvement cycle. The contribution of this work lies in combining a systematic evaluative framework with a human-centered dimension, positioning empathy as a core competency in AI ethics education. The framework is conceptual in nature but explicitly structured to guide practical implementation across diverse educational contexts, and it provides a foundation for future empirical validation through classroom pilots and cross-cultural adaptation.

Keywords

Artificial Intelligence Ethics, Empathy in AI Education, Benchmarking Ethical Competence

1. Introduction

Artificial Intelligence (AI) is no longer confined to research laboratories or niche applications; it is increasingly integrated in everyday decision-making across domains such as healthcare, finance, education, and public policy. While these systems promise efficiency and innovation, they also raise pressing concerns about fairness, accountability, transparency, and broader societal impacts. Addressing these concerns requires not only the development of technical safeguards but also the development of ethical competence among those who design, deploy, and regulate AI systems. In recent years, AI ethics education has emerged as a strategic priority. Universities, professional training programs, and policy-oriented practices are now incorporating ethics modules into their curricula (as per Table 1). However, these initiatives vary greatly in scope, depth, and effectiveness. Some programs present AI ethics as a theoretical discussion detached from practical applications; others treat it as a one-off lecture rather than an integrated theme. This heterogeneity makes it difficult to assess whether learners are acquiring the competencies needed to critically engage with the ethical dimensions of AI. The crucial challenge lies in the absence of a formal framework to evaluate the quality of AI ethics education. Without clear benchmarks, comparing programs, identifying deficiencies, and implementing targeted improvements may result in a problematic task. This paper addresses this gap by proposing a structured benchmark

2nd Workshop on Education for Artificial Intelligence (edu4AI 2025, https://edu4ai.di.unito.it/), Co-located with ECAI 2025, the 28th European Conference on Artificial Intelligence which will take place on October 26, 2025 in Bologna, Italy
*Corresponding author.
†These authors contributed equally.
enrico.barbierato@unicatt.it (E. Barbierato); alice.gatti@unicatt.it (A. Gatti); marco.gribaudo@polimi.it (M. Gribaudo)
0000-0003-1466-0248 (E. Barbierato); 0009-0008-8422-8024 (A. Gatti); 0000-0002-1415-5287 (M. Gribaudo)

University / Program	AI Literacy Initiative (non-STEM/general ed.)
Penn State (Arts & Architecture)	General-education AI literacy course open to all; covers AI basics and creative applications. [link]
SUNY System	Fall 2026: AI ethics/literacy incorporated into Information Literacy general education requirement. [link]
Ohio State University	Required General Education Launch Seminar introduces generative AI basics to undergrads. [link]
UT Austin	"Essentials of AI for Life and Society" seminar (1→3 credits), open to all students, staff, community. [link]
University of New Orleans	Free AI literacy micro-credential for all students; non-technical content about societal impact and ethics. [link]
Marist University (SUNY)	Applied AI minor open to all majors—focuses on how generative AI works and its societal effects. [link]
UAlbany (SUNY)	Artificial Intelligence and Society College & AI Plus initiative integrates AI into social sciences and humanities ethically. [link]
Sri Balaji University, Pune	Liberal Arts curriculum with embedded AI modules targeted at creative disciplines. [link]
Symbiosis International University, Pune	SAII: Interdisciplinary undergraduate AI programs for non-CS students across domains. [link]

Table 1
Universities offering AI literacy courses accessible to non-STEM or general-education students

for AI ethics education. While this work introduces a conceptual benchmark, its ultimate value lies in guiding practice. In this respect, we anticipate the need to adapt the framework to different institutional and cultural contexts, and to provide concrete implementation pathways, from undergraduate courses to professional training settings. The benchmark consists of three components: i) a set of criteria for comprehensive and effective instruction; ii) an analysis of common fallibility scenarios in which these programs fail to meet their objectives, and iii) a set of corrective actions to address these shortcomings. This work aims, by framing AI ethics education in terms of measurable standards, failure modes, and remediation strategies, to provide educators, policymakers, and accreditation bodies with a practical tool for evaluation and improvement. Furthermore, the proposed benchmark is designed to be adaptable across institutional contexts and responsive to the evolving ethical challenges posed by AI technologies. The remainder of the paper is structured as follows. Section 2 reviews the main contributions in the literature. Section 3 introduces the proposed benchmark criteria and explains their rationale. Building on this foundation, Section 4 analyzes common failure scenarios in AI ethics education, while Section 5 outlines the corrective actions designed to address them within an iterative improvement cycle. Finally, Section 6 summarizes the contributions of this work and highlights directions for future research and application.

2. Related Work

Several international organizations have issued high-level guidelines that establish foundational principles for responsible AI development and deployment. UNESCO's Recommendation on the Ethics of Artificial Intelligence [1] provides a globally endorsed framework emphasizing human rights, fairness, accountability, and inclusivity. Similarly, the Organization for Economic Cooperation and Development's (OECD) Recommendation of the Council on Artificial Intelligence [2] outlines principles for trustworthy AI, emphasizing transparency, robustness, and human-centered values. From a legal and regulatory perspective, the Ethics Guidelines for Trustworthy AI [3] outline concrete requirements, including human oversight, technical robustness, and data governance, thereby providing a bridge between abstract principles and enforceable standards. Alongside these policy frameworks, academic

initiatives have planned to embed ethical reasoning directly into AI and computer science curricula. The Embedded Ethics model [4] integrates ethical analysis throughout technical courses rather than isolating it in standalone modules, thereby fostering continuous engagement with moral questions in context. In the computing education community, practical approaches to integrating ethics have been proposed and tested. Fiesler *et al.* [5] describe methods for incorporating ethical discussions into introductory programming classes, using accessible examples and interactive activities to encourage reflection from the earliest stages of technical training. More recently, Smith *et al.* [6] have examined barriers and enabling factors in the adoption of ethics education across computing courses, gathering insights from educators on institutional support, resource availability, and assessment practices. Despite the existence of these guidelines and pedagogical experiments, there remains a notable gap: few works propose a structured and measurable framework for evaluating the quality of AI ethics education across diverse institutional and cultural contexts. The present paper addresses this gap by moving from descriptive or prescriptive accounts of "what should be taught" toward a benchmarking approach, defining explicit criteria, identifying common fallibility scenarios, and proposing corrective actions for continuous improvement.

3. Benchmark Criteria

The proposed benchmark identifies measurable dimensions that an AI ethics education program should meet to be considered comprehensive and effective. The criteria are adaptable to different institutional contexts while remaining aligned with widely recognized principles of ethical AI development.

First, programs must ensure broad content coverage. Core topics include bias and fairness, transparency and explainability, accountability and governance, privacy and data protection, and the socio-economic and cultural impacts of AI. Addressing these areas guarantees exposure to the full range of ethical challenges.

Equally important is the pedagogical approach. Ethics should be integrated across curricula, combining computer science with law, philosophy, and social sciences. Strategies such as case-based teaching, role-playing, and debates encourage students to engage actively with trade-offs, reflecting the principle of *constructive alignment* [7], which links learning activities directly to intended outcomes.

Practical application is also critical. Students should acquire hands-on skills such as detecting bias in datasets, assessing model interpretability, and conducting ethical impact assessments before deployment. These competencies align with higher-order stages of Bloom's taxonomy [8], moving from recall toward analysis, evaluation, and creation.

Another distinctive element is empathy and perspective-taking. Ethical reasoning requires anticipating the experiences of those affected by AI systems, a capacity supported by neuroscientific research on mirror neurons [9]. Structured exercises such as storytelling or perspective-switching make abstract principles more tangible and highlight the role of affective learning alongside cognitive skills.

Finally, assessment methods must reflect the multidimensional nature of ethical competence. Beyond factual recall, they should evaluate reasoning under uncertainty, the ability to integrate diverse perspectives, and the design of technically feasible yet ethically sound solutions. Following Wiggins and McTighe's *backward design* framework [10], assessments must be aligned with intended learning outcomes.

In practice, these criteria can be operationalized within existing curricular structures. Empathy-building may be introduced through short seminars or debates requiring minimal time, while bias detection can be embedded in data analysis labs. Such interventions are feasible under resource constraints and can serve as pilot modules for empirical validation.

4. Fallibility Scenarios

The following scenarios illustrate how neglecting or poorly implementing benchmark criteria can lead to superficial understanding, inadequate skill transfer, or flawed ethical judgment. By linking them to

Criterion	Description	Operational Indicators
Content Coverage	Bias, transparency, accountability, privacy, socio-economic impact.	Modules address each topic; students apply related concepts.
Pedagogical Diversity	Integration across disciplines; case-based and experiential learning.	Use of debates, role-play, or cross-disciplinary case studies.
Practical Application	Skills for bias detection, interpretability, and impact assessment.	Assignments on dataset bias or ethical audits.
Empathy and Perspective-Taking	Exercises fostering perspective-taking, informed by neuroscience.	Activities such as stakeholder storytelling or perspective-switching.
Assessment Rigor	Evaluation of reasoning, stakeholder analysis, and ethical solution design.	Open-ended case analysis, not only factual recall.

Table 2
Proposed benchmark criteria for AI ethics education, with operational indicators

real-world contexts, their relevance for educators and learners becomes clear.

A first weakness is the superficial treatment of ethics, when topics are covered in a single lecture or isolated module. Students may memorize definitions of fairness or accountability without applying them in practice, producing "checkbox ethics"—compliance in form but not in substance. This reflects findings in computing education, where ethics is often treated as an "add-on" to technical curricula, leaving students with terminology but limited reasoning skills.

A second weakness is the overemphasis on theory. Without practical exercises, students may understand bias conceptually but fail to detect or measure it in real data. Obermeyer et al. [11] showed how a U.S. healthcare risk algorithm underestimated the needs of Black patients because cost was used as a proxy for health status, demonstrating why abstract definitions are insufficient without testing and analysis.

A third weakness is the presence of cultural blind spots. Ethical principles are not culturally neutral, and frameworks designed in one region may not transfer elsewhere. UNESCO's Recommendation on AI Ethics [1] stresses the need for contextual interpretation, underscoring that curricula should include diverse case studies and, where possible, international collaboration.

Another recurring weakness is the empathy gap. Without perspective-taking, students may default to utilitarian reasoning, as illustrated in moral thought experiments such as the trolley problem [12]. Similar patterns appear in engineering education [13, 14] and in AI contexts like automated triage, where efficiency may override dignity or fairness. Structured empathy-building activities are needed to counter this tendency.

Finally, there is assessment mismatch. Traditional exams test factual recall but overlook the ability to reason under uncertainty and balance stakeholder perspectives. Wiggins and McTighe's *Understanding by Design* [10] emphasizes that authentic assessment must align with higher-order objectives. A student may excel on multiple-choice questions about privacy principles yet fail to recognize a real data breach because the scenario does not match textbook examples.

5. Corrective Actions

Addressing the fallibility scenarios described above requires a structured and systematic approach. The proposed benchmark treats improvement as an *iterative* process in which evaluation leads to targeted interventions, followed by re-assessment and further refinement. This continuous loop ensures that programs do not simply meet minimum standards once, but evolve in response to changing technological, societal, and pedagogical conditions. Table 3 summarizes the mapping between the identified fallibility scenarios and the corresponding corrective actions proposed in this benchmark, thereby providing a clear operational link between observed weaknesses and remedial strategies.

The corrective actions outlined above should be embedded within an iterative *Plan–Do–Check–Act*

Scenario	Case Example	Corrective Action
Superficial ethics	Ethics presented as an isolated lecture; observed in computing curricula [5, 6].	Integrate ethical discussions across the curriculum, embedding case analysis and reflective exercises.
Overemphasis on theory	U.S. healthcare risk algorithm underestimated needs of Black patients due to cost proxy [11].	Use experiential learning (bias detection with real datasets, simulations).
Cultural blind spots	UNESCO AI Ethics Recommendation notes need for contextualized principles [1].	Diversify case studies; include cross-cultural guest speakers and comparative analysis.
Empathy gap	Moral reasoning studies in engineering show utilitarian bias without perspective-taking ([14]).	Implement empathy-building activities such as stakeholder storytelling and role inversion.
Assessment mismatch	Recall-based exams fail to measure ethical competence; see Wiggins & McTighe [10].	Shift to competency-based evaluation using open-ended, ambiguous scenarios.

Table 3
Fallibility scenarios with documented cases and corresponding corrective actions.

(PDCA) cycle. In the planning stage, instructors define learning objectives and align them with the benchmark criteria, while also identifying potential risk areas emerging from past evaluations. The subsequent delivery phase implements the curriculum with the corrective measures already mapped to known fallibility scenarios. Evaluation then follows, combining quantitative approaches, such as competency-based assessments, with qualitative insights derived from reflective essays, peer reviews, and direct feedback from both learners and instructors. The final stage of the cycle consists of acting on these results by refining teaching methods, updating case studies, and adjusting assignments or assessment tools to address the gaps that have been identified. Beyond conceptual mapping, the next step involves empirical validation. Pilot studies could be conducted in diverse institutions—such as short-term elective modules in computer science programs or continuing education workshops for professionals—to assess feasibility, cultural adaptability, and measurable learning outcomes. Longitudinal evaluation of these pilots would provide evidence for refining the benchmark into a validated tool. Although conceptually straightforward, the practical application of this cycle is influenced by institutional resources, faculty expertise, and the flexibility of curricula. Large-scale changes may not always be feasible immediately. A pragmatic strategy involves introducing pilot modules or elective courses in which corrective actions can be tested and refined before being scaled across the curriculum. Parallel to this, faculty development workshops play a crucial role in equipping instructors with the necessary skills to teach ethical reasoning and design empathy-building activities. To further strengthen cultural inclusivity, open educational resources and collaborative networks can be leveraged to share diverse case studies and innovative assessment instruments.

6. Conclusion

This study has advanced the discussion on how to evaluate ethics education in Artificial Intelligence by proposing a benchmark that brings together content, pedagogy, application, and assessment within a unified structure. Unlike existing initiatives, which often remain descriptive or principle-based, the framework is explicitly oriented toward evaluation and continuous improvement, offering a means to identify shortcomings and suggest remedies.

At present the proposal is conceptual, but it has been crafted with implementation in mind. Its strength lies not in prescribing a universal curriculum, but in offering criteria that can be adapted to the constraints and opportunities of different institutions. The next step is empirical: pilot projects are needed to test how the benchmark functions in practice, whether it can be scaled across diverse cultural and disciplinary contexts, and how it might be refined through evidence gathered from classrooms and

training programs.

The longer-term ambition is for this benchmark to evolve into a reference tool that educators, policymakers, and accrediting bodies can draw upon to gauge the quality of AI ethics instruction. By linking measurable standards to mechanisms for improvement, it provides a pathway for transforming ethics from a peripheral concern into an integral and assessable dimension of technical education. In doing so, it contributes to shaping a generation of practitioners and decision-makers who are better prepared to confront the ethical challenges posed by intelligent technologies.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] UNESCO, Recommendation on the ethics of Artificial Intelligence, https://unesdoc.unesco.org/ark:/48223/pf0000381137, 2021. Adopted by the General Conference on 23 Nov 2021.

[2] OECD, Recommendation of the Council on Artificial Intelligence, OECD Legal Instruments, OECD/LEGAL/0449, 2019. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449, adopted on 22 May 2019.

[3] M. Cannarsa, Ethics Guidelines for Trustworthy AI, in: The Cambridge Handbook of Lawyering in the Digital Age, Cambridge University Press, Cambridge, 2021, pp. 283–297.

[4] B. J. Grosz, D. G. Grant, K. Vredenburgh, J. Behrends, L. Hu, A. Simmons, J. Waldo, Embedded ethics: Integrating ethics broadly across computer science education, Communications of the ACM 62 (2019) 54–61. doi:10.1145/3330794.

[5] C. Fiesler, M. Friske, N. Garrett, F. Muzny, J. J. Smith, J. Zietz, Integrating ethics into introductory programming classes, in: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education (SIGCSE '21), ACM, New York, NY, USA, 2021, pp. 1027–1033. doi:10.1145/3408877.3432510.

[6] J. J. Smith, B. H. Payne, S. Klassen, D. T. Doyle, C. Fiesler, Incorporating Ethics in Computing Courses: Barriers, Support, and Perspectives from Educators, in: Proceedings of the 54th ACM Technical Symposium on Computer Science Education V.1 (SIGCSE '23), ACM, New York, NY, USA, 2023, pp. 367–373.

[7] J. Biggs, Enhancing teaching through constructive alignment, Higher Education 32 (1996) 347–364. doi:10.1007/BF00138871.

[8] L. W. Anderson, D. R. Krathwohl, P. W. Airasian, K. A. Cruikshank, R. E. Mayer, P. R. Pintrich, J. Raths, M. C. Wittrock, A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Longman, New York, 2001.

[9] G. Rizzolatti, C. Sinigaglia, The mirror mechanism: a basic principle of brain function, Nature Reviews Neuroscience 17 (2016) 757–765. doi:10.1038/nrn.2016.135.

[10] G. Wiggins, J. McTighe, Understanding by Design, expanded 2nd ed., Association for Supervision and Curriculum Development (ASCD), Alexandria, VA, 2005.

[11] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations, Science 366 (2019) 447–453. doi:10.1126/science.aax2342.

[12] J. J. Thomson, The trolley problem, The Yale Law Journal 94 (1985) 1395–1415.

[13] L. L. Bucciarelli, Ethics and engineering education, in: N. A. of Engineering (Ed.), Emerging Technologies and Ethical Issues in Engineering, National Academies Press, Washington, DC, 2008, pp. 91–101.

[14] A. Colby, W. M. Sullivan, Ethics teaching in undergraduate engineering education, in: N. A. of Engineering (Ed.), Emerging Technologies and Ethical Issues in Engineering, National Academies Press, Washington, DC, 2008, pp. 27–34.