

# TraspAgente: A Scenario-Based Educational Framework for AI Transparency and Responsible Decision-Making

Giacomo Balduzzi<sup>1,\*</sup>, Manuel Striani<sup>2,\*</sup>

<sup>1</sup>Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali (DIGSPES), Università del Piemonte Orientale, Alessandria, Italy

<sup>2</sup>Dipartimento di Scienze ed Innovazione Tecnologica (DiSIT), Università del Piemonte Orientale, Alessandria, Italy

## Abstract

Generative Artificial Intelligence (GenAI) is increasingly being used by students for writing, information seeking, translation, image generation, and interaction with conversational agents. This diffusion poses a major challenge for Computer Science education: students must understand not only how AI systems work, but also when their outputs should be accepted, verified, revised, or rejected. Several existing studies share the belief that AI has necessitated a radical rethinking of educational models related to the use and relationship of technologies. The authors propose an AI literacy approach grounded in the premise that the diverse applications of artificial intelligence are not merely technological innovations to be learned and used, but also new models and forms of social change inextricably intertwined with technological evolution. Understanding and governing such changes requires multidisciplinary, interdisciplinary, and transdisciplinary theories, tools, and approaches.

Following this perspective, the paper presents *TraspAgente* (TranspAgent), a scenario-based educational framework for teaching AI transparency as responsible and permission-aware decision-making in lower secondary education. This framework combines open-ended AI-use scenarios, a structured transparency sheet, collaborative debate, and iterative revision. It translates key concepts, including reliability, bias, authorship, privacy, accountability, transparency, and permission, into observable learning artefacts and a reproducible evaluation protocol. The paper describes the framework's theoretical foundations, instructional workflow, implementation materials, and assessment design. As a design-stage contribution, it does not claim empirical learning gains, but offers a replicable basis for future classroom studies on AI literacy and responsible AI use.

## Keywords

Artificial Intelligence Literacy, Computer Science Education, Critical Decision-Making, Digital Citizenship, Scenario-Based Learning

## 1. Introduction

Generative Artificial Intelligence (AI) systems are becoming part of ordinary school practice through conversational agents, AI-enhanced search engines, writing assistants and image generators. Their accessibility creates new opportunities for learning support, but it also introduces a specific challenge for Computer Science education: students may use AI outputs as if they were neutral, verified and authoritative, while such outputs are produced by probabilistic, data-driven systems that can be inaccurate, biased, incomplete or inappropriate for the task at hand.

This challenge cannot be reduced to a problem of tool usage or school discipline. When students use AI systems for homework, research, communication or personal advice, they make decisions about reliability, source verification, originality, authorship, privacy, bias and responsibility. These dimensions are central to contemporary Computer Science because AI systems are socio-technical systems: they combine computational models, data, interfaces, institutional contexts and human interpretation. Accordingly, AI education should help learners understand both the technical nature of AI systems and the conditions under which their outputs may responsibly inform human decisions.

---

AI-LEAP 2026 - 1st International Workshop on AI Learning, Education And Practice, co-located with AIxIA 2026, University of Perugia, October 2026

\*Corresponding author.

<sup>†</sup>These authors contributed equally.

✉ giacomo.balduzzi@uniupo.it (G. Balduzzi); manuel.striani@uniupo.it (M. Striani)

ORCID 0000-0003-4448-1515 (G. Balduzzi); 0000-0002-7600-576X (M. Striani)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The increasing pervasiveness of AI across social life and virtually every dimension of human existence, from everyday routines to activities in both personal and professional contexts, has been accompanied by growing interest among scholars and decision-makers in AI education and AI literacy.

Current debates on AI in education often oscillate between two positions. One position treats AI as a technical subject, focusing on data, models, algorithms and prediction. The other treats AI as an ethical or civic issue, focusing on risks, rights and regulation. Both perspectives are necessary, but they are not sufficient unless they are connected through classroom practices that make learners' reasoning visible. In particular, younger students cannot be expected to inspect model internals or perform formal algorithmic audits. They can, however, learn to ask transparency-oriented questions: What is the source of this output? What evidence supports it? Which risks or biases may be present? What should be checked before using it? Who remains responsible for the final decision?

Public and societal debates on AI often divide between techno-optimistic and technophobic positions, reflecting a recurrent pattern in the history of technological innovation. Between these two opposing positions lies a more balanced perspective: AI is neither good nor bad; it is simply a new tool that we need to learn to use well. This argument is perhaps the most persuasive and is widely reflected in the literature and conference discussions on AI literacy. However, it also risks oversimplifying the issue. AI is far from a wholly new phenomenon: scholarly discussions on the subject can be traced back to 1950, the year in which Alan Turing's pioneering paper was published. In recent years, increased computing capacity and major technological breakthroughs have enabled new AI technologies, such as large language models, which are distinguished by their pervasive presence across work-related and broader societal contexts.

These technologies have profoundly transformed human-technology interaction, creating situations in which people may not even be aware that they are interacting with AI systems [1]. In other words, the "just-a-tool assumption" frames AI literacy primarily as a competency-based construct [2, 3]. By contrast, we argue that AI should not be understood merely as a new technological tool, but rather as a profound social transformation. AI systems are, in fact, socio-technical systems, wherein the social and technological components are neither causes nor consequences of each other, but rather shape one another through ongoing dynamic interaction. Such a holistic conception elevates AI literacy beyond a mere set of skills by embedding the technology it refers to within its broader social context. Rather than a catalogue of skills, however broad or interdisciplinary, we define AI literacy as a form of proficiency. Building on a definition originally developed for digital literacy, we conceive this proficiency as enabling people "to appropriately respond, in socially recognized ways, to even future challenges through sharing and creating knowledge and eventually participating in society" [4]. From this perspective, AI literacy constitutes a form of social and civic competence, empowering individuals to navigate societal shifts, exercise critical ethical judgment, and make informed, reflective, and responsible decisions.

In line with this framework, the present paper presents *TraspAgente*, a scenario-based educational framework that operationalises AI transparency as responsible decision-making. The framework is designed for lower secondary education but can be adapted to other school levels. Its central idea is that transparency should not be taught only as an abstract principle or as a technical property of a model. Instead, it should be enacted through a structured workflow in which students analyse realistic AI-use scenarios, choose a course of action, document their criteria, discuss alternative viewpoints and revise their decision when appropriate.

The paper makes the following contributions:

- C1.** It defines an educational interpretation of AI transparency as learner-centred responsible decision-making.
- C2.** It proposes a reproducible scenario-based workflow for AI literacy and responsible AI education.
- C3.** It specifies classroom artefacts, implementation options and an evaluation protocol for analysing students' transparent reasoning about AI use.
- C4.** It positions the framework within Computer Science education by connecting AI concepts, human-centred AI, digital citizenship and classroom assessment.

The remainder of the paper is organised as follows. Section 2 discusses related work and research positioning. Section 3 presents the *TraspAgente* framework. Section 4 describes the evaluation protocol and expected outputs. Section 5 discusses implications and threats to validity. Section 6 concludes the paper.

## 2. Background and Related Work

This section positions *TraspAgente* within the Computer Science education literature on AI literacy, human-centred transparency, generative AI education and responsible AI learning. Rather than providing a broad survey of AI in education, the review focuses on work that is directly relevant to the paper’s methodological contribution: transforming abstract principles such as transparency, trustworthiness, reliability and accountability into classroom practices that produce observable student artefacts.

### 2.1. AI Literacy and Human-Centred Transparency

As outlined above, AI literacy denotes the set of competencies and proficiencies required to understand, use, evaluate, and critically reflect on AI systems, enabling informed and ethical decision-making [5, 2, 3]. In Computer Science education, this includes conceptual knowledge about data, models, prediction, uncertainty and automation, together with the ability to recognise that AI systems are shaped by design choices, training data, interfaces and contexts of use.

As highlighted in the aforementioned review in [1], the three main phenomena underlying what we have defined as the “social transformation” brought about by AI are: 1) autonomy; 2) learning; and 3) inscrutability. AI technologies are autonomous, as they can act without human intervention. These decisions can have far-reaching consequences for individuals and society as a whole, often without people being aware that they have been influenced by AI systems. Learning represents a key feature of contemporary AI systems: through data-driven processes, they can adapt their performance and generate increasingly sophisticated outputs. Recent advances in big data have further strengthened these learning capabilities, enabling AI systems to support complex decision-making processes involving audio, text, and object recognition, as well as generative capacities. Inscrutability is one of the central challenges associated with contemporary AI development. As generative algorithmic models become increasingly complex, both their internal processes and their outputs may be interpretable only by a restricted group of experts or may remain opaque even to human observers. The growing sophistication of AI technologies allows developers to construct systems characterised by millions of interconnected parameters, whose decision-making processes frequently exceed the limits of human understanding. The distinctive features of AI highlighted by the three facets discussed above pose unprecedented challenges for human–technology interaction, with profound implications for people as moral and social agents, and for society as a whole. To face such challenges, AI literacy should not be reduced to operating AI tools or memorising technical terminology; it should also involve evaluating outputs, understanding limitations and identifying the boundary between automated suggestions and human responsibility.

Transparency is a closely related concept in AI research, where it is commonly associated with explainability, interpretability and accountability [3, 6]. Explainable AI provides technical methods for making model behaviour more understandable, while human-centred AI investigates how people interact with intelligent systems and how design decisions influence trust, reliance and decision-making [7, 8]. These areas are central to trustworthy AI, but they do not directly solve the educational problem addressed in this paper. Lower secondary students cannot be expected to inspect model architectures, training distributions or feature attributions. However, they can learn to make their own interaction with AI systems transparent by documenting sources, assumptions, risks, checks and responsibilities.

Recent educational frameworks move in this direction by adapting responsible and trustworthy AI concepts to learning contexts. Balduzzi et al. propose a hybrid educational framework for AI ethics in healthcare that combines an e-learning platform with LLM-based interaction to support medical students in analysing ethical issues such as informed consent, bias, safety, transparency, equity and legal

regulation [9]. TeenTrust-AI similarly translates the ALTAI principles of trustworthy AI into a structured educational framework for teenagers, using teaching, learning and trustworthiness verification phases to guide students in assessing an AI-powered system through checklist-based activities [10]. These contributions show the relevance of structured learning designs for responsible AI education, but they focus respectively on domain-specific ethics training in healthcare and on the verification of AI trustworthiness against a predefined set of principles.

*TraspAgente* complements these approaches by focusing specifically on transparency as a decision-making competence. It does not propose a new XAI algorithm or an AI tutoring architecture. Instead, it defines a classroom workflow in which the main analytical object is the decision trace produced by students when they decide whether, how and why an AI output should be used. This positioning is relevant for Computer Science education because it connects computational understanding, human-centred AI and responsible user agency.

## 2.2. Generative AI Education, Digital Citizenship and Scenario-Based Learning

Generative AI introduces educational risks that are both technical and social. Generated outputs may be fluent but inaccurate, biased, unoriginal or unsafe to disclose [11]. In classroom activities, these risks affect research, writing, translation, image production and chatbot interaction. They include hallucinated information, superficial learning, uncritical delegation, plagiarism, unclear authorship, copyright concerns and privacy violations. Since generative AI systems generate outputs through statistical modelling rather than truth verification, students need explicit strategies for evaluating plausibility, evidence and context of use.

Responsible AI and digital competence frameworks emphasise fairness, transparency, human oversight, accountability, information literacy, safety and fundamental rights [12, 13, 14, 15]. Critical studies of data and automation further show that algorithmic systems may reproduce exclusion, bias and inequality [16, 17, 18, 19]. These perspectives imply that AI education should not isolate technical content from civic responsibility. Students should learn that AI systems are embedded in social relations and that their use may affect privacy, authorship, representation and participation.

Scenario-based and collaborative learning are appropriate for this purpose because responsible AI use rarely has a single correct answer. Open-ended scenarios expose learners to ambiguity, competing values and trade-offs, while group discussion supports argumentation, perspective taking and revision. Previous AI education work has already shown the usefulness of structured scenarios and guided questions for ethics-oriented and trustworthiness-oriented learning, for example in healthcare ethics education [9] and in teenager-centred trustworthy AI verification [10]. *TraspAgente* builds on this pedagogical logic and adapts the APE civic-education method [20] to generative AI use in school contexts. The APE project was established to investigate how formal and informal learning could be integrated within civic education. To this end, it designed and implemented the educational workshop that laid the foundations for *TraspAgente*. The workshop engaged more than 660 students from Italian upper secondary schools across multiple regions and educational tracks, including technical, scientific, and humanities programs.

By simulating individual and collective decisions about the use of AI, it encourages critical reflection and fosters open, nuanced discussion. In doing so, it makes explicit what often remains implicit: every use of AI involves a decision, and every decision carries trade-offs in terms of benefits, costs, and risks.

The distinctive contribution of *TraspAgente* is to move from principles to a reproducible classroom workflow. Students do not only discuss whether AI is trustworthy or ethical in general terms; they analyse a concrete scenario, choose a course of action, complete a transparency sheet, defend their reasoning in debate and revise their decision when appropriate. The framework therefore turns transparency into an assessable educational artefact and provides a methodological bridge between AI literacy, digital citizenship and Computer Science education practice.

Building on the methodology of the APE project, *TraspAgente* tailors its approach to a new target audience: lower secondary school students (aged 11–14). Beyond developing scenarios centered on the ethical dilemmas of AI in education, the project adapts its language and interaction style specifically for

this age group. Case study content and phrasing will be co-designed with teachers from participating schools, who will collaborate directly with researchers by contributing practical insights into real-world AI use in learning environments. By blending the expertise of a multidisciplinary research team (spanning computer science, sociology, and law) with the field experience of educators across various subjects, the *TraspAgente* framework embodies a fully transdisciplinary approach to workshop design [21].

### 3. The *TraspAgente* Framework

*TraspAgente* is a design-oriented framework for teaching AI transparency through classroom decision-making. The framework is based on four design requirements. First, the activity must involve authentic AI-use dilemmas rather than decontextualised definitions. Second, students must be required to justify their decisions through explicit criteria. Third, peer discussion must expose students to counterarguments and alternative viewpoints. Fourth, the activity must generate artefacts that teachers and researchers can analyse without relying on opaque automatic scoring.

#### 3.1. Workflow Framework

Figure 1 summarises the operational workflow of *TraspAgente*. The figure should be read from left to right as a learning cycle that transforms AI use from an implicit individual action into an explicit, discussable and revisable decision process.

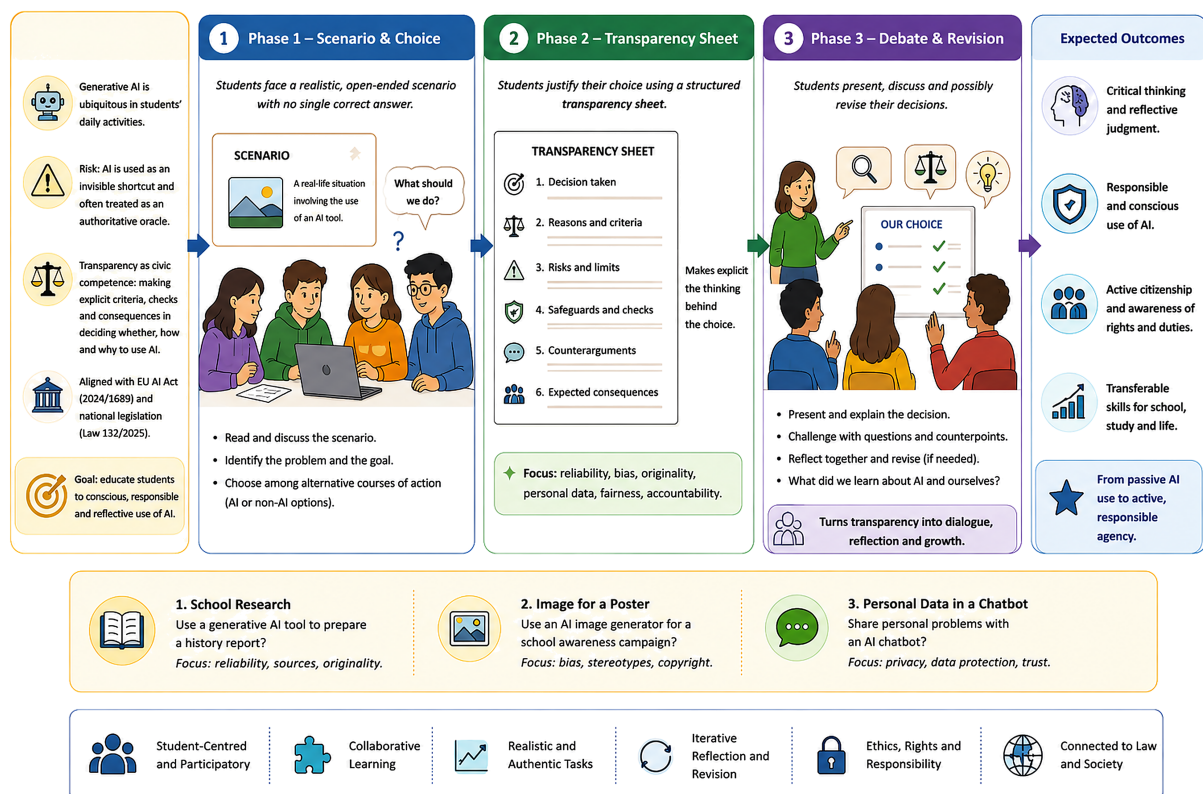


**Figure 1:** The *TraspAgente* framework for scenario-based AI transparency, literacy and critical thinking.

The left part of Figure 1 defines the educational problem. Generative AI is increasingly present in students' everyday activities, but it may be used as an invisible shortcut or treated as an authoritative oracle. The framework therefore introduces transparency as a civic and computational competence: students learn to make explicit the criteria, checks and consequences involved in deciding whether, how and why an AI tool should be used.

**Table 1**

Initial *TraspAgente* scenario repository and main transparency dimensions.

| Scenario                   | Decision dilemma                                           | Transparency focus                                            |
|----------------------------|------------------------------------------------------------|---------------------------------------------------------------|
| School research            | Use a generative AI tool to prepare a history report.      | Reliability, source verification, originality and authorship. |
| Image for a poster         | Use an AI image generator for a school awareness campaign. | Bias, stereotypes, copyright and representation.              |
| Personal data in a chatbot | Share personal problems with an AI chatbot.                | Privacy, data protection, trust and safeguards.               |

The central part of the figure describes the three-phase workflow. In **Phase 1**, students analyse a realistic, open-ended scenario involving an AI tool and choose among alternative courses of action, including AI-based and non-AI-based options. In **Phase 2**, they complete a transparency sheet that asks them to document the decision taken, reasons and criteria, risks and limits, safeguards and checks, counterarguments and expected consequences. In **Phase 3**, students present their choices, compare them with other viewpoints and revise their decision when debate reveals weaknesses, missing evidence or better safeguards. The workflow is recursive: the goal is not only to select an option, but to improve the quality and accountability of the reasoning behind that option.

The lower part of Figure 1 shows how the workflow is instantiated through concrete scenarios, such as using generative AI for school research, producing an image for a poster, or sharing personal data with a chatbot. Each scenario foregrounds a specific set of transparency concerns: reliability and source verification, bias and stereotypes, copyright and originality, privacy and trust. The final strip of the figure summarises the pedagogical principles of the framework: student-centred participation, collaborative learning, realistic tasks, iterative reflection, ethical responsibility and connection with law and society.

Overall, the workflow turns transparency into a decision trace. For each scenario, students produce evidence of what they decided, why they decided it, which risks they identified, which safeguards they proposed and whether discussion changed their reasoning. This trace is the main educational and analytical artefact of the framework.

### 3.2. Classroom Artefacts and Implementation

The framework uses four artefacts: scenario cards, transparency sheets, debate prompts and teacher observation notes. Scenario cards describe a realistic school problem and a small set of possible actions. Transparency sheets collect the initial decision, reasons, criteria, risks, safeguards, counterarguments, expected consequences and revised decision. Debate prompts support structured peer discussion. Teacher observation notes record facilitation issues, misconceptions and participation patterns.

The initial scenario repository is reported in Table 1. The scenarios were selected because they represent frequent student interactions with generative AI and cover different dimensions of responsible AI use.

Table 2 summarises the transparency sheet. These fields also define the coding dimensions for future empirical analysis.

*TraspAgente* is intentionally designed as a low-infrastructure educational system. It can be implemented with printed materials, in a hybrid format, or through a learning-management system. The reference implementation is paper-based because it is easy to replicate and avoids unnecessary processing of minors' data. A future digital implementation may support anonymised data export, rubric-based coding and teacher dashboards, but the core educational mechanism remains the classroom discussion around explicit decision traces.

**Table 2**

Transparency sheet fields and analytical purpose.

| Field                 | Analytical purpose                                         |
|-----------------------|------------------------------------------------------------|
| Decision taken        | Identifies the selected course of action.                  |
| Reasons and criteria  | Captures the explicit basis of the decision.               |
| Risks and limits      | Captures awareness of uncertainty, harm or misuse.         |
| Safeguards and checks | Captures verification strategies and protective actions.   |
| Counterarguments      | Captures exposure to alternative viewpoints.               |
| Expected consequences | Captures responsibility and anticipation of outcomes.      |
| Revised decision      | Measures whether discussion changes or enriches reasoning. |

## 4. Evaluation Protocol and Expected Outputs

The present paper reports the design stage of the framework. For this reason, it does not present empirical results or claim measured learning gains. Instead, it specifies an evaluation protocol that can be used in future classroom studies and identifies the expected outputs of the intervention.

### 4.1. Study Design

A suitable evaluation design is a mixed-methods pre/post study with lower secondary school students. Before the activity, students complete a short questionnaire on AI use, perceived reliability and decision criteria. During the workshop, groups complete transparency sheets for one or more scenarios. After the debate and revision phase, students complete a post-questionnaire and a final reflection activity. When feasible, a comparison group can receive a standard AI-awareness lesson without the transparency sheet and debate cycle. This comparison would allow researchers to evaluate whether the structured workflow adds value beyond general awareness raising.

The primary data sources are completed transparency sheets, teacher observation notes, debate summaries or transcripts, and pre/post questionnaires. Ethical and data-protection requirements are central because the target participants are minors. Data collection should use anonymous identifiers, avoid unnecessary personal data and comply with the General Data Protection Regulation (GDPR) [22]. In scenarios concerning chatbots and personal information, students should reason about hypothetical cases rather than disclose private experiences to real systems.

### 4.2. Metrics and Analysis

The evaluation should combine qualitative coding with descriptive statistics. Qualitative analysis should identify which transparency dimensions appear in student responses and how specific, coherent and critical their justifications are. Quantitative indicators can summarise the frequency of dimensions, the number of revised decisions, pre/post changes in questionnaire responses and inter-rater agreement among coders. Table 3 reports the core evaluation indicators.

The expected outputs of the intervention are not automatic scores but analysable educational artefacts. Each transparency sheet records a decision trace that can be used for formative assessment and research. At the group level, the scenario repository enables comparison across different AI-use contexts. At the study level, the protocol supports analysis of whether students move from generic statements such as “AI can be wrong” to more specific reasoning involving source verification, data minimisation, authorship, safeguards and accountability.

For reproducibility, future studies should report the number of classes and students, session duration, number of scenarios, group size, teacher training procedure, whether students used real AI tools or pre-generated examples, questionnaire items, rubric definitions, coding manual and anonymisation procedure.

**Table 3**

Core evaluation indicators for future classroom studies.

| Indicator                  | Evidence source                | Interpretation                                                         |
|----------------------------|--------------------------------|------------------------------------------------------------------------|
| Transparency coverage      | Transparency sheet coding      | Number and diversity of transparency dimensions explicitly considered. |
| Justification quality      | Rubric-based scoring           | Depth, specificity and coherence of reasons, criteria and safeguards.  |
| Revision index             | Initial and revised decisions  | Whether debate leads to changed decisions or enriched reasoning.       |
| Reliability awareness      | Questionnaire and sheet fields | Ability to distinguish plausibility from verified information.         |
| Bias and privacy awareness | Scenario-specific responses    | Recognition of stereotypes, data risks and protective strategies.      |
| Inter-rater agreement      | Independent coding             | Reproducibility of the qualitative analysis, e.g., Cohen's $\kappa$ .  |
| Teacher feasibility        | Observation notes and feedback | Practical usability, timing and classroom manageability.               |

## 5. Discussion and Threats to Validity

*TraspAgente* contributes to Computer Science education by conceptualising AI literacy as a social capacity enacted through accountable decision-making, rather than reducing it to conceptual knowledge about algorithms or to general ethical awareness. This distinction is important because students increasingly encounter AI systems through everyday interfaces before they encounter formal instruction about machine learning. A framework for AI education must therefore connect technical concepts, such as data-driven generation and uncertainty, with user-level practices, such as checking evidence, recognising bias, protecting personal data and taking responsibility for the final use of an output.

Furthermore, the concept behind the *TraspAgente* workshop—both in its overall framework and core methodological premises—strongly echoes the key principles of Edgar Morin's ecology of action [23]. The late sociologist argued that all actions, including moral ones, tend to escape the control of their authors as they enter a web of inter-retroactions within the environment in which they take place. Furthermore, every action confronts the ultimate unpredictability of its consequences. Actors suffer from a systematic ignorance regarding the outcomes of their actions and can only know the effects of a decision a-posteriori. Morin also highlighted certain dilemmas inherent in action choices—such as the tension between risk and precaution, both of which remain necessary despite their opposition. Another dilemma is about the means-ends relationship: means can corrupt and cause the ends to degenerate. One of the key strengths of the *TraspAgente* workshop lies in its capacity to help students recognise and reflect on the complexity of the choices, consequences, and dilemmas that arise from the everyday use of AI systems.

Compared with lecture-based AI awareness, *TraspAgente* produces visible artefacts that can be discussed, assessed and revised. Compared with purely technical AI activities, it foregrounds the human decision process surrounding AI use. Compared with principle-based ethics lessons, it offers a reproducible classroom workflow. This makes the framework suitable for integration into Computer Science, civic education and interdisciplinary digital literacy curricula.

The main limitation is that the framework is currently design-oriented. Empirical evidence is needed to determine whether it improves students' AI literacy, whether the transparency sheet captures meaningful reasoning, and whether the approach is feasible across diverse school contexts. Internal validity may be affected by teacher facilitation, group dynamics and novelty effects. External validity is limited because the scenarios are designed for lower secondary education and may require adaptation for other ages, countries or curricula. Construct validity is also a concern: written transparency sheets measure articulated reasoning, not necessarily long-term behaviour or internal understanding. Finally, reproducibility depends on the release of scenario cards, transparency sheet templates, coding manuals

and analysis procedures.

These limitations define the next stage of the research. Future classroom studies should include multiple classes, common teacher guides, independent coders and, where feasible, comparison conditions. Digital implementations should be evaluated carefully: they may improve data collection, but they should not replace the peer discussion and revision processes that are central to the framework.

## 6. Conclusion and Future Work

This paper presented *TraspAgente*, a scenario-based educational framework for teaching AI transparency as responsible decision-making. The framework addresses a timely problem in Computer Science education: students increasingly use generative AI systems, but they often lack structured practices for evaluating reliability, authorship, bias, privacy and accountability. *TraspAgente* responds by combining realistic scenarios, a transparency sheet, collaborative debate and iterative revision.

The contribution is methodological and design-oriented. Rather than proposing a new AI model or an automated assessment system, the paper defines a reproducible educational workflow that makes students' reasoning about AI visible and analysable. The framework links AI literacy, human-centred AI, digital citizenship and classroom assessment, supporting a shift from passive AI use to reflective and accountable agency.

Future work will implement the workshop in lower secondary classrooms, collect transparency sheets and debate evidence, refine the coding rubric, estimate inter-rater agreement and compare the framework with less structured AI-awareness activities. Additional scenarios will address AI-assisted translation, recommendation systems, synthetic media and automated feedback. A digital version may also be developed to support anonymised data collection and teacher feedback while preserving the central role of collaborative discussion.

## Declaration on Generative AI

During the preparation of this work, the authors used Claude Opus 5 in order to paraphrase, reword, and improve writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

- [1] M. Pinski, A. Benlian, Ai literacy for users – a comprehensive review and future research directions of learning methods, components, and effects, *Computers in Human Behavior: Artificial Humans* 2 (2024) 100062. URL: <https://www.sciencedirect.com/science/article/pii/S2949882124000227>. doi:<https://doi.org/10.1016/j.chbah.2024.100062>.
- [2] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, M. S. Qiao, Conceptualizing ai literacy: An exploratory review, *Computers and Education: Artificial Intelligence* 2 (2021) 100041. doi:[10.1016/j.caeai.2021.100041](https://doi.org/10.1016/j.caeai.2021.100041).
- [3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, *Information Fusion* 58 (2020) 82–115. URL: <https://www.sciencedirect.com/science/article/pii/S1566253519308103>. doi:<https://doi.org/10.1016/j.inffus.2019.12.012>.
- [4] J. K. Njenga, Digital literacy: The quest of an inclusive definition, *Reading & Writing* 9 (2018) a183. URL: <https://doi.org/10.4102/rw.v9i1.183>. doi:[10.4102/rw.v9i1.183](https://doi.org/10.4102/rw.v9i1.183), published 30 August 2018.
- [5] D. Long, B. Magerko, What is ai literacy? competencies and design considerations, in: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–16. doi:[10.1145/3313831.3376727](https://doi.org/10.1145/3313831.3376727).
- [6] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017. URL: <https://arxiv.org/abs/1702.08608>. arXiv:1702.08608.
- [7] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Su, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for human-ai interaction, in: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–13. doi:[10.1145/3290605.3300233](https://doi.org/10.1145/3290605.3300233).
- [8] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe and trustworthy, *International Journal of Human–Computer Interaction* 36 (2020) 495–504. doi:[10.1080/10447318.2020.1741118](https://doi.org/10.1080/10447318.2020.1741118).
- [9] G. Balduzzi, T. Balduzzi, M. Striani, Design an hybrid educational framework for AI ethics in healthcare: Leveraging LLMS and e-learning platforms to empower medical students (full paper), in: E. Marengo, M. Ponticorvo, M. Striani (Eds.), *Proceedings of the 1st International Workshop on Education for Artificial Intelligence co-located with the 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024)*, Bozen-Bolzano, Italy, November 27, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024, pp. 116–129.
- [10] S. Mitrović, M. Striani, F. Flammini, Teentrust-AI: A structured educational framework for empowering teenagers to evaluate trustworthiness of ai, in: E. Marengo, M. Ponticorvo, M. Striani (Eds.), *Proceedings of the 2nd Workshop on Education for Artificial Intelligence (edu4AI 2025)*, co-located with the 28th European Conference on Artificial Intelligence (ECAI 2025), CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2025, pp. 112–121. Published 2025-12-01.
- [11] UNESCO, Guidance for Generative AI in Education and Research, United Nations Educational, Scientific and Cultural Organization, Paris, 2023. URL: <https://unesdoc.unesco.org/ark:/48223/pf0000386693>.
- [12] R. Vuorikari, S. Kluzer, Y. Punie, DigComp 2.2: The Digital Competence Framework for Citizens: With New Examples of Knowledge, Skills and Attitudes, Technical Report, Publications Office of the European Union, Luxembourg, 2022. doi:[10.2760/115376](https://doi.org/10.2760/115376).
- [13] L. Floridi, J. Cowsls, A unified framework of five principles for ai in society, *Harvard Data Science Review* 1 (2019). doi:[10.1162/99608f92.8cd550d1](https://doi.org/10.1162/99608f92.8cd550d1).
- [14] 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>.
- [15] European Parliament and Council of the European Union, Regulation (eu) 2024/1689 of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act), Official

- Journal of the European Union, 2024. URL: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
- [16] C. O’Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown, New York, NY, USA, 2016.
  - [17] S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York University Press, New York, NY, USA, 2018.
  - [18] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, St. Martin’s Press, New York, NY, USA, 2018.
  - [19] M. Broussard, *Artificial Unintelligence: How Computers Misunderstand the World*, MIT Press, Cambridge, MA, USA, 2018.
  - [20] G. Balduzzi, collaborators, *Ape method for civic education: Scenario-based deliberation and participatory learning*, Project materials and educational documentation, 2024.
  - [21] T. S. Goetze, Integrating ethics into computer science education: Multi-, inter-, and transdisciplinary approaches, in: *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1, SIGCSE 2023*, Association for Computing Machinery, New York, NY, USA, 2023, p. 645–651. URL: <https://doi.org/10.1145/3545945.3569792>. doi:10.1145/3545945.3569792.
  - [22] European Parliament and Council of the European Union, Regulation (eu) 2016/679 of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation), *Official Journal of the European Union*, 2016. URL: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>.
  - [23] E. Morin, *La Méthode, 6 : Éthique*, Éditions du Seuil, Paris, 2004. URL: <https://www.seuil.com/ouvrage/la-methode-6-edgar-morin/9782020786386>, UNESCO catalogue record: <https://unesdoc.unesco.org/ark:/48223/pf0000137647>.