12th International Conference of the Immersive Learning Research Network
iLEAD Proceedings iLRN2026

DOI: https://doi.org/10.56198/qyqs8b64

AI Detectives 101: A Digital Escape-Room Workshop for Critical AI Literacies, Co-Creation, and Ethical AI Design

Yvoni Galetaki[1] and Iro Voulgari[1]

[1]University of Malta, Institute of Digital Games, Msida, Malta
yvonnigaletaki@gmail.com, iro.voulgari@um.edu.mt

Abstract. This work presents an immersive escape-room workshop designed to foster critical AI literacies, co-creation, and ethical reflection. The pilot study involved 13 participants who collaboratively engaged in puzzles that simulate how AI systems are trained, how bias emerges, and how design decisions shape inclusivity and fairness. Guided by Justice, Equity, Diversity, and Inclusion (JEDI) principles, and using a game-based format, the workshop aimed to engage participants with AI concepts through interactive challenges that encourage reflection on the ethical design of technology. This study explores three research questions: (1) how the escape room format enhances a digital game to promote critical AI literacies and ethical reflection, (2) how JEDI framing in both the art medium and AI system design influences ethical reasoning, and (3) how co-creation tasks foster participants' sense of agency in shaping inclusive AI systems. Findings showed that while participants struggled with factual recall in individual tasks, they demonstrated greater engagement, collaboration, and ethical reasoning in open-ended activities. This progression from bias detection to inclusive design thinking illustrates the potential of the escape room format to foster not only active learning and motivation but also value-sensitive design practices. Participants' generation of diverse solutions showcase how plural perspectives enrich the redesign of ethical AI systems.

Keywords: AI Literacy, Escape Rooms, Digital Games Based Learning, Immersive Learning, Co-Creation, Design Thinking.

Introduction

Artificial intelligence (AI) is introducing powerful tools and new opportunities for education, prompting educators to reconsider established approaches to teaching and learning, while posing critical ethical challenges [1]. This work introduces an activity titled "*AI Detectives 101",* developed as an extension of the ArtBot digital game through an escape room scenario. This activity is designed to foster critical AI literacies, co-creation, and ethical reflection [2]. By leveraging the principles of Justice, Equity, Diversity, and Inclusion (JEDI), the workshop bridges technical understanding with ethical critique, allowing participants to interrogate algorithmic processes, identify biases, and redesign systems for inclusivity and fairness. The integration of AI into education has sparked global discourse on its ethical implications, as highlighted by Chen [3] and in UNESCO's guidelines on AI ethics [4]. Ethical concerns such as algorithmic bias, data privacy, and human agency underscore the need for educational interventions that promote AI literacy [2, 3]. Research on AI literacy emphasizes the importance of equipping learners with competencies to critically evaluate AI technologies, understand their societal impacts, and engage in value-sensitive design processes [3]. Similarly, studies on immersive learning environments demonstrate their potential to transform abstract concepts into tangible, collaborative experiences, fostering deeper engagement and critical thinking [5]. Building on these foundations, the "*AI Detectives 101"* workshop employs an escape room format to situate participants in a narrative where they act as investigators examining real-world AI challenges, such as bias in training data and algorithmic decision-making. This approach aligns with findings that immersive play enhances motivation, collaboration, and ethical reflection [5]. By integrating the JEDI principles into both the art medium (e.g., painting, sculpture, music or video) and AI system design the workshop positions participants as active co-designers rather than passive consumers.

To implement this activity, the workshop builds on ArtBot [6], an AI education-focused digital game developed within the LearnML project [7], which aimed to support AI literacy among primary and secondary education students. Through ArtBot, participants explore cultural content and the ethical practices of AI design, confronting biases and reflecting on the design of systems for inclusivity. Within a cultural heritage narrative, players train and guide an AI helper, and explore concepts such as supervised and reinforcement learning, data labeling, model parameters, and algorithmic decision-making. Its design support constructivist learning through hands-on experience, while its cultural heritage theme enables interdisciplinary use across subjects such as computing, arts, history and humanities.

Building on these ideas, this paper presents the design of the "*AI Detectives 101"* workshop as an escape room aimed at strengthening critical AI literacies. Through a JEDI-informed co-design approach, participants move from identifying bias to evaluating outcomes by reflecting ethically, fostering a deeper understanding of AI systems' behavior [8]. The embodied and creative interactions can make abstract AI concepts more accessible, supporting learners to connect to familiar experiences [9]. As a result, this hands-on activity aims to lead to the responsible use of AI technologies, and enhance understanding of the societal impact [8]. In this work, we define critical AI literacies as the ability to: (i) explain core AI concepts, (ii) identify where bias and limitations enter socio-technical systems, and (iii) reason about consequences and mitigation strategies in context. The main questions guiding our study were:

RQ1: How does the escape-room format promote critical AI literacies and fosters ethical reflection?

RQ2: How does the integration of JEDI principles in both the art medium and AI system design influence participants' ethical reflections?

RQ3: How does co-creating narratives or features within the escape room foster participants' sense of agency in shaping more inclusive AI systems?

Methodology

Escape Room Design

The digital escape room was implemented using the Nearpod platform [11], an interactive teaching platform for educators allowing them to design game-based, teacher-guided lessons for maximizing motivation and engagement. The modules offered by Nearpod featured real-time quizzes, drawing and collaborative boards, interactive videos, polls, and virtual simulations enhancing immersion. For an escape-room approach, these modules allow to sequence puzzles as "missions" through the different features and progressively reveal core AI-related ideas like training data, patterns, bias, and evaluation - while providing the facilitator with real-time visibility into responses to scaffold, adapt, and debrief. The design of the escape room puzzles utilized Bloom's taxonomy scaffolding [10] to guide participants from foundational knowledge to higher-order reflection and creation. It spanned 60-90 minutes and was organized into five sequential activity phases, each aligned with specific cognitive processes and framed through JEDI principles:

- **Remember/Understand** - *Detective 101:* Participants begin by playing the first level of ArtBot, where they are introduced to basic AI concepts and the notion of algorithmic bias. Through short prompts and an embedded quiz, participants recall and recognize examples of biased outcomes in AI systems (Fig. 1). This phase establishes a baseline understanding, while JEDI framing encourages participants to consider whose perspectives are represented or omitted in the system's behavior.
- **Understand/Apply** - *Algorithm Anatomy*: Building on this foundation, participants identify the core components of supervised learning, including data collection, labeling, training, and prediction. Through guided mapping activities, they apply these concepts by reconstructing the training pipeline of the game's AI system. The JEDI lens highlights how assumptions embedded in data selection and labeling practices can reinforce inequities when transferred into real-world contexts.
- **Analyze** - *Impact Check*: In this phase, participants analyze how differences in training datasets influence model predictions across diverse communities. Using impact maps (Fig. 2) and scenario-based prompts, they compare outcomes for different user groups and examine patterns of harm, exclusion, or misrepresentation (Fig. 2). This analytical step shifts attention from how AI works to whom it affects, foregrounding systemic bias and social consequences.

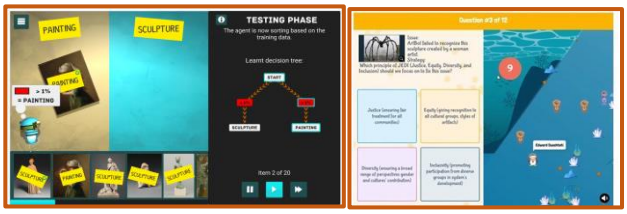

Fig. 1. From left to right: Playing the ArtBot game; taking a quiz on Nearpod platform.

Fig. 2. From left to right: Scenario-based prompt cards; Anatomy of an Algorithm; mind maps and collaborative boards.

- **Evaluate** - *Priorities Check*: Participants critically evaluate alternative AI design features by weighing trade-offs such as accuracy, transparency, fairness, and accessibility (Fig. 2). Priority grids and structured discussion prompts support value-based judgment, explicitly grounded in JEDI principles. Participants justify their choices and reflect on how design priorities shape ethical responsibility in AI systems.
- **Create** - *AI Innovators*: The final phase invites participants to synthesize their learning by co-creating inclusive AI solutions. Working collaboratively, they design new models, narratives, or game features that address previously identified biases. These creations are documented on shared collaborative boards, emphasizing inclusive design thinking, plural perspectives, and creative problem-solving.

The workshop was conducted via Zoom, with participants assigned to breakout rooms in small groups to support different roles and paces. Facilitators provided instructions verbally and in written form, using clear visuals and simple language to accommodate varying levels of prior AI knowledge. Participants collaborated through shared visual artifacts across all phases and submitted their solutions via Google Form.

Participants

This pilot study involved 13 participants, 6 boys and 7 girls, English-speakers. The mean age (in years) of the participants was M=15, except for one participant who was a 35-year-old educator. Participants brought diverse prior experiences with digital games and AI-related concepts. To encourage collaboration and exposure to diverse perspectives, participants were randomly assigned into pairs, ensuring that most collaborators were unfamiliar with one another. This pairing strategy was intended to reduce bias in group formation and to promote equitable participation through shared problem-solving and dialogue across different age groups and experience levels.

Data Collection and Analysis

We followed a qualitative approach and collected qualitative data through a) observation notes by the facilitator focused on engagement, collaboration patterns, understanding, b) the generated digital artifacts and the collaborative outputs, and c) the log files and response rates recorded by the Nearpod platform. All data collected during the workshop were fully anonymous. Participants contributed using code names. No personal information was stored.

Findings

Analysis of the *AI Detectives 101* session revealed contrasting patterns across activity types. Overall participation rates, based on the platform rates, were moderate (50%), which appears partly attributable to the workshop structure: participants began the first activity individually and only later transitioned into teams. Performance on the multiple-choice items assessing understanding was moderate overall with an average score of 67%, indicating that factual recall and understanding of AI concepts (e.g., supervised learning steps, training data, prediction)

remained challenging at the point where participants were still working independently, despite a short introduction and initial gameplay. In contrast, engagement was notably stronger in open-ended and creative tasks, especially the *Draw It* and *Collaborate Board* activities, where participants were discussing their ideas to develop an appropriate strategy. Participants contributed visualizations, explanations, redesign ideas, and scenario-based reflections (Fig. 4) more consistently in these formats.

Fig. 4. Scenario reflections and algorithmic anatomy.

Contribution patterns were uneven; several participants produced high-quality, substantial individual input, while others contributed minimally, suggesting heterogeneous levels of AI literacy and confidence. The collaborative boards nonetheless showed early evidence of a shift from bias detection toward design-oriented thinking, with participants proposing inclusive redesigns even when solutions were partial or incomplete (Fig. 5).

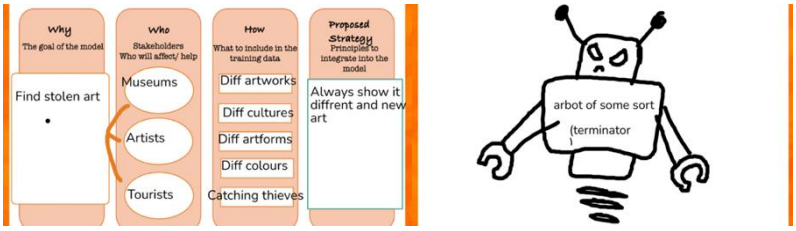

Fig. 5. System re-design.

These patterns speak directly to RQ1. Lower performance on individual multiple-choice items suggests that brief instruction and gameplay alone do not automatically translate into precise procedural understanding. However, the richer reasoning observed in interactive, puzzle-like activities (drawing, shared boards, collective problem-solving) indicates that the escape-room structure was doing its intended work: it created a context where participants externalized thinking, built meaning through interaction, and connected AI mechanisms to real-world consequences supporting the integrative reasoning central to critical AI literacies. In relation to RQ3, collaborative artifacts show participants moving beyond bias detection toward proposing concrete modifications, indicating an emerging designer stance ("how we would redesign"). Uneven contributions caution, however, that co-creation does not automatically produce equal shared ownership for all learners. The overall shift toward inclusive redesigns, grounded in JEDI principles, provides preliminary support for RQ2.

Discussion and Conclusions

This pilot study involved a small and diverse group of participants, enabling exploration and testing of the escape-room activities. Participants were randomly assigned into pairs to promote collaboration and exposure to different perspectives. However, random pairing did not control for differences in prior knowledge, confidence, or familiarity with collaborative tools, which likely contributed to uneven participation across activities.

This study offers exploratory insights into engagement, co-creation, and ethical reasoning, as they emerged in this workshop. Certainly, the findings cannot be generalized or interpreted as causal evidence of learning outcomes, but they provide valuable lessons and can inform directions for future work. The escape-room format successfully enhanced engagement and collaboration, particularly in open-ended tasks such as Draw It and Collaborate Board. These activities encouraged participants to actively contribute ideas, visualizations, and redesign proposals, demonstrating integrative reasoning and creative sense-making. Participants were able to connect technical concepts to social implications through shared artifacts and discussion.

However, the findings also revealed challenges in technical understanding, as participants struggled with factual recall and procedural knowledge during independent activities. This suggests that while the escape-room format motivated participation and fostered collaboration, additional or differently timed scaffolding is needed to support technical precision alongside creative exploration. The format's immersive, narrative-driven design aligns

with prior research on game-based learning, which emphasizes interactivity, presence, and social negotiation as key drivers of deeper learning outcomes.

The integration of JEDI principles was evident in participants' progression from bias detection toward inclusive design thinking during collaborative tasks. In particular, the Collaborative Board activity highlighted participants' ability to critique biased assumptions and propose redesigns that incorporated fairness, accessibility, and inclusivity, even when solutions remained partial or exploratory. This pattern aligns with the hypothesis that explicitly embedding JEDI scaffolds supports ethically grounded reasoning. At the same time, heterogeneous AI literacy levels among participants point to the need for differentiated support to ensure equitable participation and sustained ethical reflection across groups. These findings underscore the importance of integrating JEDI principles not only as discussion prompts but as design constraints embedded within both the digital artifacts and the AI system itself. Co-creation tasks were among the most engaging components of the workshop and demonstrated participants' sense of ownership and agency. The observed shift from identifying bias to proposing inclusive features and narratives suggests that co-creation can serve as a mechanism for fostering agency and empowering participants to envision themselves as capable of shaping more inclusive AI systems.

Overall, the pilot showed the potential of combining immersive escape-room formats, JEDI principles, and co-creative design activities to foster critical AI literacies and ethical reflection. Future iterations should refine scaffolding strategies, particularly to better support technical understanding. This approach offers a promising pathway for culturally aware, participatory, and design-oriented AI education.

Acknowledgments

The authors would like to thank the organizers of the Science in the City Festival for providing the opportunity to design and run this workshop. They would also like to acknowledge the support of the Institute of Digital Games at the University of Malta since this workshop was delivered under its initiative. They are grateful to all participants for their enthusiasm, creativity, and thoughtful engagement throughout the activity, which greatly enriched the workshop experience. This work was supported by the Pulse-Art Horizon Europe project, funded by the European Union's Horizon Europe research and innovation programme under Grant Agreement No. 101178988.

References

1. du Boulay, B.: Artificial Intelligence in Education and Ethics. In: Handbook of Open, Distance and Digital Education. pp. 93–108. Springer, Singapore (2023). https://doi.org/10.1007/978-981-19-2080-6_6.
2. Long, D., & Magerko, B. What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–16). ACM (2020).
3. Chen, H.: The ethical challenges of educational artificial intelligence and coping measures: A discussion in the context of the 2024 World Digital Education Conference. Science Insights Education Frontiers 20(2), 3263–3281 (2024). https://doi.org/10.15354/sief.24.re339
4. UNESCO: Recommendation on the Ethics of Artificial Intelligence. UNESCO, Paris (2021). https://unesdoc.unesco.org/ark:/48223/pf0000380455
5. Radianti, Jaziar & Majchrzak, Tim A. & Fromm, Jennifer & Wohlgenannt, Isabell. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. Computers & Education. 147 (2020). 10.1016/j.compedu.2019.103778
6. ArtBot Game., Website: https://art-bot.net (https://art-bot.net/)
7. Voulgari, Iro & Zammit, Marvin & Stouraitis, Elias & Liapis, Antonios & Yannakakis, Georgios. Learn to Machine Learn: Designing a Game Based Approach for Teaching Machine Learning to Primary and Secondary Education Students. 593-598 (2021). 10.1145/3459990.3465176.
8. Chiu, T. K. F., Ahmad, Z., Ismailov, M., & Sanusi, I. T., What are artificial intelligence literacy and competency? A comprehensive framework to support them. Computers and Education Open (2024).
9. Darabipourshiraz, H., Rollins, S., Trajkova, M., Zhang, C., Belghith, Y., McKlin, T., Roberts, J., Magerko, B., & Long, D., Designing and evaluating museum exhibit prototypes to foster middle schoolers' AI literacy through creativity and embodiment. Twentieth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '26), March 08–11, 2026, Chicago, IL, USA (2026).
10. Krathwohl, David R. "A revision of Bloom's taxonomy: An overview." Theory into practice 41.4: 212-218, (2002).
11. Mallinson, Daniel J. "Nearpod: Classroom Learning Platform, (https://nearpod. com)." Journal of Political Science Education18.3: 411-416 (2022).