



WP3: Participatory Action Research on Needs and Prioritisation Map



TABLE OF CONTENTS

Chapter 4 Ethical concerns	2
4.1 FATE (Fairness, Accountability, Transparency, Ethics) Principles	2
4.2 Ethical Challenges in AI	3
4.2.1 Bias and Fairness	3
4.2.2 Transparency and Explainability.....	4
4.2.3 Accountability and Responsibility	5
4.2.4 Data Privacy and Protection	6
4.3 Hypothetical Case Studies.....	6
4.3.1 Example 1: An AI-Powered Tutoring System.....	6
4.3.2 Example 2: Classroom Management AI Monitoring Student Engagement.....	7
4.4 Fairness Paradox	7
4.4.1 Fairness Paradox Illustrated with Examples from Hypothetical Case Studies.....	8
4.4.2 Strategies for Educators and Policymakers	8





Chapter 4 Ethical concerns

As technological innovations permeate the landscape of higher education, implementing Artificial Intelligence (AI) systems offers extraordinary opportunities and significant ethical considerations. Ethical AI in education refers to the thoughtful and principled creation and application of AI technologies that respect all participants' values, needs, and rights within the educational sphere. The focus is not only on making AI effective, but also on ensuring it is used in a morally sound and socially responsible way.

The importance of ethics in AI for education cannot be overstated. It shapes how AI systems are designed, who they are designed by, and how they impact students and teachers. Ethically aligned AI helps build trust among users, enhances the educational experience, and ensures that advancements in AI contribute positively to educational outcomes without causing harm or perpetuating inequities.

Ethical AI is vital for education because it stands at the intersection of technology's limitless possibilities and the intrinsic value of educational integrity and equity. Education systems shape knowledge and skills and influence societal norms and individual behaviors. Unethical AI in these settings can perpetuate biases, reduce transparency, and challenge the fairness of educational opportunities, thereby reinforcing existing societal disparities. For instance, an AI system that is not meticulously designed might inaccurately predict student success based on biased data, potentially discriminating against underrepresented student groups. Alternatively, AI systems might prioritize certain learning styles over others, disadvantaging students who do not fit a predetermined mold. Similarly, without stringent ethical controls, AI could compromise student data privacy, exposing vulnerable populations to exploitation or manipulation. Thus, advocating for and establishing ethical AI practices in education is not just a necessary regulation—it is imperative to ensure these technologies serve as tools of empowerment rather than tools of exclusion.

4.1 FATE (Fairness, Accountability, Transparency, Ethics) Principles

To effectively contextualize and implement ethical AI in education, it is imperative to delve into the FATE principles that guide the establishment of principled AI frameworks. These principles act as pillars ensuring that AI systems in educational settings do not merely function efficiently but do so in a manner that aligns with safeguarding human interests and upholding democratic values. Each component of FATE addresses critical ethical considerations, grounding these systems in practices that promote inclusivity, responsibly trace decision-making, create transparency, and prioritize ethical standards.

Fairness: AI systems in education must be free from biases that can disadvantage any group of users. Fair treatment for all users is crucial to ensure the technology is just and equitable.





Accountability: There should be mechanisms to hold designers and operators of AI systems responsible for how these systems operate. Recognizing that the development of AI technologies is not neutral supports proactive measures to address potential issues and rectify them transparently.

Transparency: Openness in AI processes and decisions crucially aids understandability and trustworthiness. Stakeholders should know how and why AI systems are making decisions in educational settings.

Ethics: AI's alignment with ethical values ensures that these technologies act consistently with the common good, respecting the privacy, dignity, and rights of individuals and demonstrating sensitivity to the complexities of educational environments.

4.2 Ethical Challenges in AI

4.2.1 Bias and Fairness

Bias in AI arises when an AI system behaves prejudicially due to flawed assumptions in its algorithm or data inputs. In educational contexts, if the data used to train an AI system encompasses historical biases or unequal representation of various groups, the AI might amplify these biases instead of overcoming them. For example, an AI tutoring system might be more effective for students from certain backgrounds if the training data predominantly represents those populations, potentially deepening educational disparities rather than closing gaps. Overcoming this challenge necessitates stringent auditing processes, diverse data sets, and continuous adjustments to ensure AI assesses and interacts with all students equitably.

Fairness in AI within an educational setting encompasses a lack of prejudice and equal treatment across diverse student demographics. However, achieving fairness can be immensely challenging as biases from various sources can infiltrate AI systems. For instance, if a dataset used for training an AI for college admissions primarily includes information from historically affluent schools, it may unintentionally favor applicants from similar backgrounds. Similarly, language processing tools might perform inadequately for non-native English speakers if primarily trained on datasets dominated by native speech, potentially resulting in lower evaluations of their written assignments. These scenarios stem from what is often societal bias--preconceptions that are ingrained in societal structures being unwittingly encoded into AI systems. Such unfairness can significantly impact students' educational opportunities and experiences, wherein certain groups might be systematically and disproportionately disadvantaged. Identifying and continually correcting these embedded biases is crucial to fortifying the fairness of AI in education.





Historical and societal biases deeply embedded in our culture can inadvertently seep into AI systems, resulting in unfairly biased machine learning (ML) models. When an AI system leverages historical data to make predictions or decisions, it essentially draws on a dataset that may already reflect past inequalities or stereotypes. For instance, if historical academic performance data show disparities among different demographics (based on socio-economic status, ethnicity, etc.), using such data directly to train ML models can lead to these systems perpetuating or accentuating the disparities they embody. The outcomes or predictions made by these biased systems then have real-life implications—they can influence university admissions, scholarship allocations, or even career advice—affecting the opportunities and treatment of individuals based on flawed or predisposed assessments. Hence, these decisions reinforce the existing societal norms by continually favoring those already advantaged while potentially disenfranchising others, thereby feeding back into the societal data cycle with new data tainted by the same old biases. This looping effect—a self-fulfilling prophecy of bias sustenance—only deepens divides unless active, continuous interventions interrupt this chain, spot skewness early, and reconfigure processes to more equitable grounds, ensuring AI systems serve to bridge rather than widen gaps born from historical and societal biases.

4.2.2 Transparency and Explainability

When AI systems operate in education, the decisions they make can significantly influence students' academic journeys and professionals' teaching strategies. However, AI algorithms' inner workings, often called 'black boxes,' can be opaque. The lack of transparency makes it difficult for users, including students, educators, and administrators, to understand how or why a decision was made. This can erode trust in AI applications and potentially hinder their effectiveness and acceptance.

Explainability ties directly to the ethical call for transparency, referring to the ability of AI systems to justify the decisions or recommendations they make in understandable terms. In educational settings, where decisions can affect career trajectories and personal growth, stakeholders should be aware of AI decisions and understand the rationale behind them. The priority should be promoting insights into AI processes, which helps nurture confidence among users, facilitate acceptance, and enable stakeholders to correct or refine AI systems over time realistically and collaboratively.

Explainable AI (XAI) refers to techniques and practices designed to help humans understand and trust the outputs provided by AI systems. This is crucial in education, where decisions influenced by AI can impact educational trajectories significantly. Recent approaches in XAI incorporate layer-wise relevance propagation, which visually indicates how neural networks make decisions, and local interpretable model-agnostic explanations (LIME), which help to clarify predictions by approximating the underlying model locally with an interpretable one.





Another important technique is SHAP (SHapley Additive exPlanations), which assigns each feature an importance value for a particular prediction. By integrating XAI techniques like these into the development and deployment of AI systems in higher education, institutions can ensure that decisions affecting students' academic trajectories are transparent, interpretable, and trustworthy. This level of explainability allows educators and administrators to not only critique AI-driven outcomes but also make informed adjustments that benefit the entire educational process, promoting fairness and fostering trust in AI systems.

4.2.3 Accountability and Responsibility

Determining accountability in AI-driven decisions poses profound ethical challenges. Unlike traditional systems, AI's decision-making process involves complex algorithms and extensive datasets, making it challenging to pinpoint responsibilities when something goes wrong—such as issuing an unfair grade or denying educational opportunities based on skewed data processing. Addressing this issue requires clear regulatory frameworks, responsibilities assigned to both developers and users of AI tools, and thorough documentation of decision-making processes.

Accountability also involves responsive recourse where affected persons can contest unjust decisions. It necessitates a governance framework that embeds accountability within every phase of AI design and deployment, emphasizing ethical considerations from the outset. It also requires more than just ethical intention; it asks for mechanisms encouraging ethical practices, including educating AI developers on the ethical considerations within artificial intelligence and ensuring educational institutions understand their obligations when implementing AI.

Accountability for AI in education involves delineating clear responsibility for the AI system's deployment and ensuring exceptions or malfunctions are appropriately addressed. For instance, when an AI fails to predict students' learning challenges accurately and subsequently recommends insufficient or incorrect resources, it should be clear which body - the technology provider, the institution, or a specific educational department - is responsible for correcting the error. Furthermore, this accountability should extend into anticipatory planning, considering potential failures and designing AI systems that can either accommodate manual oversight or alert humans when unexpected patterns or results occur. On the governance side, educational institutions could establish an ombudsman or similar role specializing in AI-related issues to handle disputes, maintain fairness, and navigate situations where responsibility might be ambiguous. Schools and higher education entities may also draft a set of standards outlining their commitments and describing how they maintain accountability in their deployment of educational AI technologies. This structured and preemptive approach to accountability





creates an environment where AI tools can be questionably more beneficial and integrated into educational frameworks with minimal disruption and optimal efficiency.

4.2.4 Data Privacy and Protection

With AI systems continuously collecting and analyzing vast amounts of data, the potential to misuse sensitive personal information is a significant concern. If safeguards are insufficient, there could be leaks or unauthorized exploitation of private data, thus compromising the privacy and security of students and educators. Compliance with strong data protection laws like the EU GDPR is vital, ensuring data collected for educational advancements respects individual privacy rights without compromise.

Moreover, ethical considerations necessitate broader discussions beyond mere adherence to regulatory requirements. They advocate for secure design perspectives prioritizing user consent steps, minimal data retention periods, and strict limitations on data usage. They reiterate the necessity for educational AI applications to be built on principles of respect for user autonomy and privacy, fostering safer digital learning environments.

4.3 Hypothetical Case Studies

Hypothetical case studies can vividly illustrate how ethical dilemmas play out in implementing AI in education—highlighting potential risks and possible resolution strategies.

4.3.1 Example 1: An AI-Powered Tutoring System

Imagine a university rolling out an AI-powered tutoring system intended to assist students in understanding complex mathematical concepts. The data used to configure the system has been sourced from several high-performing schools with limited demographic diversity. As a result, the customized learning modules are inadvertently better suited for those demographics, inadvertently bypassing nuances that could aid different backgrounds. Over time, monitoring students' academic performance reveals that while average scores have increased from the primary demographic, there's a declining trend or stagnation in scores among students from other ethnic groups.

Upon more in-depth investigation, it is shown that the example problems built into the AI system involve contexts not familiar to everyone universally—such as scenarios highlighting golf or skiing, sports statistically favored by upper socio-economic groups—a subtlety creating a biased learning experience. Under these mounting ethical concerns, the university must now redesign the AI's algorithms, incorporating a broader dataset that reflects a diversity of experiences and social contexts. In response, educators collaborate with AI developers to





create a more inclusive model involving user input changes and expanded scenarios broadly recognizable across all student groups, which helps personalize and, therefore, improve the learning outcomes across the previously underperforming demographics.

4.3.2 Example 2: Classroom Management AI Monitoring Student Engagement

Consider a secondary school that uses a classroom management AI to monitor student engagement and participation. This system utilizes real-time data from students' interactions with digital devices, course platforms, and even physiological signs via wearable technology to gauge attention and engagement, adapting instruction methods accordingly. However, concerns about surveillance and data privacy arise. Students and parents are growing uncomfortable with the pervasive data collection—everything from the links the students click on to how long they gaze at the screen is logged and assessed.

An unintended consequence comes to light when it's noticed that the AI system, while effective in some regards, pressures students into feigning attention, thus lowering the authentic quality of engagement and making them feel perpetually monitored and anxious. To address this, the school holds an open forum for feedback and decides to implement key changes. They reassess what data is truly beneficial and necessary for educational purposes and opt to protect and restrict access according to the strictest privacy standards. This reversal toward a more mindful application of AI, focusing solely on crucial pedagogical factors while respecting student privacy, shifts the school culture back to a trust-based model, enhancing relationships and student wellbeing.

These hypothetical scenarios demonstrate complex issues around educational AI implementation, revealing that credible risks can emerge despite the best intentions. Effective response strategies are paramount—correcting biases and respecting privacy—and upholding education's role as a foundation for responsible and inclusive growth.

4.4 Fairness Paradox

The "Fairness Paradox" in AI illuminates a complex challenge: the prioritization or balancing of different fairness metrics can sometimes lead to unintended consequences (Corbett-Davies, 2023). Essentially, the process of tuning AI systems to optimize certain fairness outcomes might inadvertently aggravate other inequities. For instance, an AI algorithm adjusted to equalize false positive rates across different demographics might inadvertently lead to unbalanced false negative rates, difficulty recognizing this balancing act between different fairness metrics is vital to implementing ethical AI systems in education.

When we assess AI applications in the educational domain through an array of fairness metrics (like equality of opportunity, representation equity, etc.), we might inadvertently compromise one form of fairness while enhancing another. Thus, deploying AI solutions that





are tailored to diverse educational settings without keen oversight might unintentionally perpetuate new modalities of bias and discrimination.

4.4.1 Fairness Paradox Illustrated with Examples from Hypothetical Case Studies

Example 1 Revisited (AI-Powered Tutoring System): In the AI-powered tutoring system in a university, suppose the tool was calibrated to ensure that all students, regardless of their backgrounds, showed improvement in their scores. In striving for this metric of fairness—performance equity—the AI algorithm might enhance content delivery according to the initial performance inputs, whereby stagnating students might end up receiving slightly favored treatment with more repetitive and simpler materials adjusted to their learning. This optimization can unintentionally undermine the growth of quicker learners who may be put through the same pacing, not providing them with the challenges needed for advanced learning. Though this fairness management enables students from lower performance brackets to catch up, it might unjustly cap the progress opportunities for those who could advance more rapidly, hence disrupting the balance of fairness across varying academic capabilities.

Example 2 Revisited (Classroom Management AI Monitoring Student Engagement): The calibration of AI to ensure uniform engagement across all students could also be seen as a manifestation of the Fairness Paradox. If the classroom management AI prioritizes maintaining consistent, measurable engagement levels, it might unfairly penalize naturally less expressive or introverted students who might not outwardly display engagement but internalize and process information effectively. By misaligning the measure of engagement (visible activity versus intellectual engagement), the system could distinctly advantage the extroverted students, thereby instigating a fairness trade-off between visible engagement and cognitive absorption.

4.4.2 Strategies for Educators and Policymakers

To navigate through the fairness paradox, educators and policymakers in higher education need to adopt a multidimensional lens toward fairness. This entails continuous tweaking and real-time evaluations of AI systems to understand disparate impacts, revolving the strategies around dynamic, contextually apt definitions of fairness rather than static or over-generalized models. Having a robust feedback mechanism where students can report perceived inadequacies or biases in AI applications can also provide fresh insights into reconfiguring fairness heuristics. Furthermore, embracing multidisciplinary cooperation in developing these systems—integrating insights from pedagogy, sociology, and ethics besides technology—can synergize efforts to rectify and proactively guard against biases leading to an unfair paradox





in educational settings. This comprehensive approach can elucidate and mitigate the intricate fairness paradoxes, guiding the responsible use of AI in education.



Funded by the
European Union

HITS project has received
funding by the European Union – Italian National Agency INDIRE
Agreement n. **2023-1-IT02-KA220-HED-000152340**