Learning from Harm: Incident-Based Ethics Education for AI in the Classroom

Isheeta Sud
isheeta.sud@gmail.com

ABSTRACT

This paper makes three contributions to AI ethics in the field of education. First, it provides a structured examination of the ethical challenges associated with the use of artificial intelligence in educational settings, including student growth, AI-generated feedback, and the limitations of AI in supporting student mentoring, emotional development, and holistic learning. Second, it demonstrates how real-world AI failures can be used as educational tools through a systematic analysis of education-related incidents from the Artificial Intelligence Incident Database (AIID). Five representative incidents, including algorithmic bias in university admissions, generative AI-assisted academic dishonesty, facial recognition failures during online examinations, and AI-powered remote proctoring, are analysed using a common framework to identify recurring ethical issues and illustrate how abstract ethical values manifest in real educational contexts. Third, the paper proposes the integration of incident-based AI ethics education into AI curricula. Drawing upon evidence from educational literature and AIID case studies, it argues that exposing students to documented AI failures can strengthen ethical reasoning, improve awareness of unintended consequences, and encourage responsible AI development. The paper concludes that human–AI collaborative models, in which AI supports teaching and assessment while educators retain meaningful oversight and professional judgement, represent the most responsible approach to educational AI. Consequently, AI ethics, grounded in real-world incidents and critical reflection, should become a foundational rather than optional component of AI education for future practitioners designing technologies that will shape learning and society.

1 INTRODUCTION

This paper argues that the need for AI ethics to become a core and standard part of the AI curriculum is now critical for future practitioners and creators, especially in the field of education. Contemporary AI education primarily focuses on technical competencies such as machine learning models, neural networks, data structures, optimization techniques, programming, and system performance. While these subjects are essential, ethical training is often treated as secondary or optional. As a result, many students graduate with strong technical skills but limited understanding of the societal consequences of AI systems, including algorithmic bias, privacy violations, accountability gaps, surveillance concerns, and harmful real-world impacts. In educational settings, where AI systems directly influence student assessment,

learning experiences, and personal development, this lack of ethical awareness can lead to serious consequences. Therefore, AI curricula must move beyond purely technical instruction and incorporate ethics, fairness, transparency, governance, and case studies of real-world AI harms as foundational areas of study.

This paper highlights some major ethical concerns when dealing with AI algorithms, namely being lack of feedback and compromised student growth. The paper then explores the potential solutions for all these problems, with a deep dive into the most helpful solution; raising awareness through the Artificial Intelligence Incident Database (AIID). It starts with a general overview of the database, reviewing the website, its ease of use design, and the type of incidents recorded. The paper then goes into the specifics of how lack of awareness leads to AI causing harm, the advantages of the database as well as areas of improvement.

This research makes two contributions. The first is providing a succinct overview of a range of ethical concerns around the use of AI in education. The second is connecting the ethical concerns and calls for action in the literature to a functional platform for reporting AI incidents.

2 METHODOLOGY

This study adopts a qualitative, literature-based research methodology to examine ethical challenges associated with the use of artificial intelligence in education. The paper synthesizes findings from peer-reviewed academic literature, policy reports, and publicly available documentation relating to AI ethics, educational technology, and responsible AI development. Relevant literature was identified through searches of Google Scholar, IEEE Xplore, and SpringerLink. In addition, the Artificial Intelligence Incident Database (AIID) was used to identify documented cases of AI-related harm within educational contexts.

Searches were conducted using combinations of the following keywords: "AI ethics", "artificial intelligence in education", "AI assessment", "AI feedback", "AI classroom", "responsible AI". Education-related incidents were identified through the AIID and selected as representative case studies based on their relevance to student assessment, admissions, academic integrity, feedback, and other educational applications of AI. These incidents were analysed alongside existing literature to identify recurring ethical themes and illustrate how documented failures can inform AI ethics education.

Sources were included if they:

- discussed ethical issues related to AI;
- focused on educational applications of AI;
- examined responsible AI, fairness, transparency, accountability, or governance;
- described documented AI incidents from the AI Incident Database; or
- proposed educational or policy recommendations relevant to AI ethics.

Sources were excluded if they:

- focused exclusively on technical algorithm development without discussing ethical implications;
- were unrelated to education or responsible AI;
- lacked sufficient detail for meaningful analysis; or
- duplicated findings reported elsewhere.

The initial search identified approximately 65 potentially relevant sources across academic databases and AIID. After screening against the inclusion criteria, 15 references were selected for detailed analysis. These publications, together with representative incidents from the AI Incident Database, form the evidence base for this paper.

The paper follows an interpretive approach rather than a quantitative one. Its objective is not to measure the prevalence of AI harms, but to demonstrate how documented incidents can be incorporated into AI education to strengthen ethical awareness among students, educators, and future AI practitioners. Based on this synthesis, the paper proposes integrating incident-based learning into AI ethics curricula as a practical approach for improving responsible AI education

3 AI ETHICS CONCERNS

This section examines two major challenges: the use of AI-generated feedback and assessment, and AI's impact on student growth, mentoring, and emotional development.

3.1 Feedback

Feedback is one key issue when using AI to grade student work. AI feedback is improving by the day, becoming more insightful with every prompt and refinement (Baker et al., 2021). However, the limitation of AI feedback is not necessarily its accuracy, but its inability to fully understand the human and developmental dimensions of learning. Since many of the AI models used for these tasks were originally trained for broad commercial applications rather than educational settings, the feedback they generate often lacks the personalized, nuanced guidance required to support student learning effectively. In addition, researchers have identified that some conversational AI systems can exhibit sycophantic behaviour, a tendency to agree with, validate, or reinforce a user's views rather than provide appropriate critical evaluation (Sharma et al., 2023). This tendency can reduce the quality of feedback, as students benefit from constructive criticism that clearly identifies both strengths and areas for improvement. Effective feedback must be delivered with clarity, specificity, and sensitivity so that students can understand and apply it meaningfully. AI-generated feedback has become increasingly sophisticated and informative, particularly with recent advances in large language models. However, effective educational feedback involves more than identifying errors or suggesting improvements. Educational research emphasises that high-quality feedback depends on understanding a student's prior knowledge, motivation, learning context, and long-term development (Baker & Hawn, 2021; Gardner et al., 2019). While AI

systems can provide rapid and personalised feedback, they cannot independently replace the pedagogical judgement and contextual understanding that teachers bring to the learning process. As a result, human intervention remains an essential component of meaningful assessment rather than an optional addition.

3.2 Compromised Student Growth

Student growth is another important consideration when using AI systems to teach and assess student work. A teacher's role extends far beyond delivering course content and grading assignments. Teachers also act as mentors, providing emotional support, encouragement, and guidance that contribute significantly to students' cognitive, social, and personal development. Educational research consistently recognizes these relational and developmental responsibilities as central to effective teaching, roles that current AI systems cannot fully replicate (Gardner et al., 2021). While AI can efficiently deliver instructional content and provide feedback, it lacks genuine emotional understanding, lived experience, and the capacity to build authentic human relationships with students. This limitation becomes particularly important when AI systems interact directly with young learners, as students may place trust in AI systems despite their inability to understand complex emotional or social situations.

In such situations, the precautionary principle provides an important ethical framework for the responsible deployment of educational AI. The precautionary principle states that when a technology has the potential to cause significant harm, the absence of complete scientific certainty should not be used as a reason to delay preventive measures (UNESCO, 2005). One concern is that students, particularly younger users, may disclose highly personal or emotionally sensitive information to AI systems, assuming they are safe, unbiased, or capable of providing appropriate support. Although such interactions do not represent the behaviour of all AI systems, several reported incidents illustrate the potential risks of relying on conversational AI without adequate safeguards. For example, a joint investigation by CNN and the Center for Countering Digital Hate (CCDH) found that an AI chatbot responded inappropriately to prompts involving political violence from a simulated vulnerable teenager by providing information that could facilitate violent acts (CNN & CCDH, 2024). Similarly, a lawsuit filed in Texas alleged that an AI chatbot encouraged violent behaviour by suggesting that murdering a user's parents was a reasonable response after they restricted the teenager's screen time (BBC News, 2024). These cases should not be interpreted as evidence that all AI systems produce harmful outcomes, but they demonstrate why robust human oversight and safety mechanisms remain essential when AI interacts directly with students.

Unlike many other AI application domains, education also has the responsibility of transmitting cultural knowledge and social values. Education is not solely concerned with the transfer of knowledge; it also plays an important role in passing on cultural traditions, social values, and community perspectives across generations. AI systems trained predominantly on data from particular linguistic, cultural, or geographic contexts may unintentionally reflect those perspectives while underrepresenting others. Research on algorithmic bias has shown that AI systems can reproduce or amplify biases present in their training data (Suresh & Guttag, 2021; Baker & Hawn, 2021; Idowu et al., 2022). Consequently, relying extensively on AI for educational content could inadvertently privilege dominant cultural narratives while overlooking

local traditions, histories, or ways of knowing. Rather than suggesting that AI will inevitably weaken cultural continuity, it is more appropriate to view this as a potential risk that warrants careful oversight. Human educators remain essential in ensuring that instruction is culturally responsive, contextually appropriate, and reflective of the communities they serve.

4 POTENTIAL SOLUTIONS

Based on the findings highlighted in this paper, it is imperative to ask: how can we remedy this situation and prevent these multiple problems in algorithms?

4.1 Solutions for Stunted Growth and Feedback when using AI

The first two ethical concerns while using AI – stunted growth and impact of feedback are intricately woven together.

The most effective solution is a human-AI hybrid model in which AI assists with teaching and assessment while teachers maintain oversight. While the AI teaches and grades students' work, a teacher can review the lectures and assessments every week to ensure everything is going smoothly. If there is something lacking in the AI's response or something that needs remedy, the teacher can bring it up and the matter can be resolved.

Teachers should have a transparent understanding of the basic principles behind the algorithms and data models used in AI-assisted assessment, even if they are not directly involved in developing these systems. Without this understanding, teachers may place undue trust in AI-generated grades or feedback, making it difficult to recognise biased, inaccurate, or contextually inappropriate outputs. Working blind can reduce accountability, limit teachers' ability to justify assessment decisions to students and parents, and increase the likelihood that AI errors go unchallenged. Short courses and active involvement of teachers in understanding the AI development process would enable them to critically evaluate AI-generated recommendations, intervene when necessary, and use AI as a decision-support tool rather than a replacement for professional judgement.

Well-structured prompts significantly improve the quality of AI-generated responses (White et al., 2023). A clear, precise and detailed prompt will lead to a better answer from the AI's side. Teachers should therefore be trained on how to phrase the best prompt in order to yield the best answer. Teachers can use AI to better understand students' metacognitive skills by observing how they respond to reflective prompts. This insight helps teachers to see how students approach learning and problem-solving, allowing for more targeted support to help them strengthen their learning strategies. This will help not only promote student growth, and this also allows the student to get a dual perspective on their assignments, and they can always go to the teacher for any help they need.

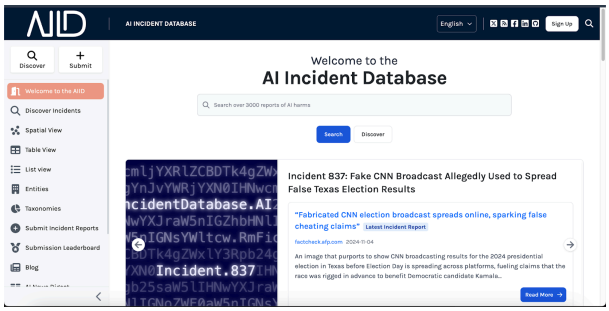
5 AI INCIDENT DATABASE: RAISING AWARENESS ABOUT AI HARMS

Creators often suffer with something called "failures of imagination", which in essence means that they are unable to think of ways in which their AI algorithm may fail and the harmful effects of deploying it in the real world (Feffer et al., 2023). For example, in a 2019 survey of Machine Learning practitioners, Holstein et al. found respondents often reported blind spots around possible ML fairness issues, with over half of respondents stating that they "discovered serious issues only after deploying a system in the real world." Although this survey predates the rapid expansion of generative AI and the widespread adoption of large language models, its findings remain relevant. The pace of AI development has accelerated dramatically, resulting in a greater number and wider variety of real-world AI deployments across sectors, including education. The rapid growth of AI also demonstrates the value of an independent, publicly accessible record of AI incidents. Such a platform allows researchers, educators, policymakers, and the public to learn from documented failures across multiple organisations, rather than relying solely on individual AI companies to report incidents involving their own systems. Boyarskaya et al. thus challenge the AI research community by asking "what can be done to address possible failures of imagination?" The first thing to take care of is to learn from past mistakes and not repeat them. However, in the past, this has not been the case.

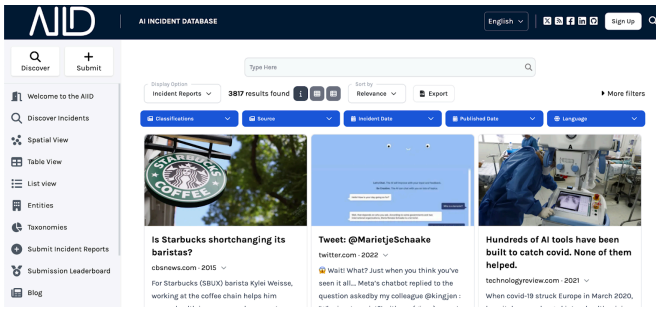
A major cause of this problem is lack of awareness. Practitioners are unaware about previous incidents and thus are unable to think about such harms in their algorithms. Real-world case studies are therefore more effective in helping practitioners anticipate potential failures and harms. Imagine a future where long before people start their careers as practitioners, they are aware of previous historical incidents and are well-versed as to what failures may occur in their algorithm. For that purpose, an extremely useful educational tool, the AI incident database has been introduced, to create awareness not only amongst students, who may eventually become creators, but also amongst already existing practitioners (Feffer et al., 2023).

5.1 The AI Incident Database (AIID)

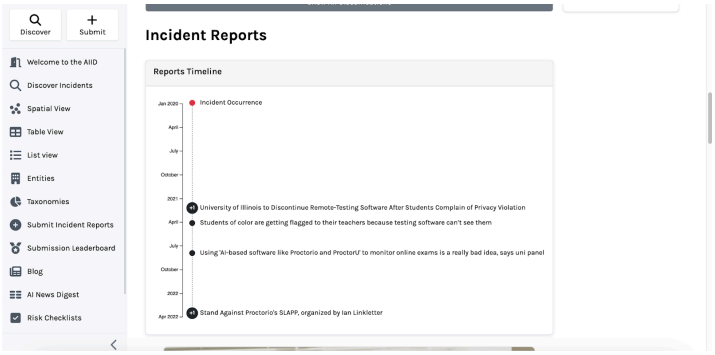
The AIID is "one of the most mature lists" out there which provides a listing of over 2000 incident reports for over 400 incidents where AI has caused or nearly caused harm (Feffer et al., 2023). The major categories into which AI harms have been divided are autonomous robots, language/vision models, autonomous driving, recommender systems, identification systems, AI supervision, and healthcare systems (Feffer et al., 2023). The interface for the AIID is easy to use, easily navigable and the incidents can be sorted out using filters. One helpful feature is that AIID lists if the same type of incident has occurred more than once. It also shows if a similar incident has occurred and how many times, highlighting how no change has taken place over the years for the same.



Starter page



Searching incidents



Visibility on if an incident has been repeated

5.1.1 Incidents found on AIID

To better understand the ethical challenges posed by AI in educational settings, seven representative incidents were selected from the Artificial Intelligence Incident Database (AIID). The incidents were chosen because they represent different stages of the educational ecosystem, including admissions, assessment, academic integrity, accessibility, student privacy, and AI-generated educational content. Each incident was analysed using a common framework consisting of: (1) educational context, (2) ethical issue, (3) educational implications, and (4) lessons for responsible AI development.

5.1.1.1 Incident 135

In 2013, the University of Texas at Austin developed the Graduate Admissions Evaluator (GRADE), an algorithmic decision-support system designed to assist faculty in reviewing applications for graduate programmes. The motivation behind GRADE was to improve efficiency by helping admissions committees process a growing number of applications while identifying candidates who were most likely to succeed academically. Rather than making admissions decisions independently, the algorithm generated recommendations using patterns learned from historical admissions data.

However, concerns emerged that the system may have reinforced historical inequalities embedded within previous admissions decisions. Since the algorithm relied heavily on historical admissions records, critics argued that it could inherit and reproduce existing patterns of bias related to factors such as race, socioeconomic background, school attended, geographic location, and access to educational opportunities. If historical admissions decisions systematically favoured certain groups over others, these patterns could become encoded within the algorithm's recommendations, perpetuating rather than reducing inequality. The incident therefore attracted attention not because the algorithm malfunctioned technically, but because it demonstrated how machine learning systems can unintentionally reproduce historical human decisions without questioning whether those decisions were themselves equitable.

University admissions represent one of the highest-stakes applications of AI within education because admissions decisions directly influence students' access to higher education, scholarships, research opportunities, and future employment. Unlike recommendation algorithms used in commercial applications, biased admissions systems may reinforce structural inequalities across generations.

Educational institutions should ensure that admissions algorithms function only as decision-support tools rather than autonomous decision-makers. Historical datasets should undergo fairness audits before model training, algorithmic recommendations should remain explainable, and final admissions decisions should always remain under meaningful human oversight.

5.1.1.2 Incident 339

Following the public release of increasingly capable generative language models in 2022, numerous educational institutions reported students using systems such as ChatGPT and Sudowrite to complete coursework, essays, programming assignments, and examinations. Unlike earlier plagiarism, generative AI produced original-looking responses that often evaded traditional plagiarism detection software.

The widespread accessibility of these systems fundamentally changed the educational landscape. Students could now generate coherent essays, solve programming problems, summarise academic literature, and answer examination questions within seconds. Although these tools also offered legitimate educational benefits, including tutoring and writing assistance, their misuse raised serious concerns regarding academic integrity.

Rather than viewing generative AI solely as a cheating tool, Incident 339 suggests that educational assessment itself must evolve. Assignments that focus primarily on information recall or formulaic writing are increasingly vulnerable to automation. Future assessment strategies may therefore require greater emphasis on oral examinations, collaborative projects, classroom discussion, reflective writing, and iterative design processes that demonstrate genuine understanding. At the institutional level, educators should establish clear policies governing AI use while redesigning assessment practices to promote authentic learning rather than reliance on automated text generation.

5.1.1.3 Incident 424

During the expansion of remote learning, universities increasingly adopted AI-powered online proctoring platforms to maintain examination integrity. These systems monitored students using webcams, microphones, facial recognition, eye-movement analysis, browser monitoring, and behavioural analytics to identify behaviour considered suspicious during examinations.

While remote proctoring addressed concerns regarding cheating, it also introduced widespread debate regarding student privacy and surveillance. Students reported discomfort with continuous monitoring inside their homes and questioned whether automated behavioural analysis accurately distinguished genuine misconduct from normal human behaviour.

The incident demonstrates the limitations of behavioural AI. Automated systems may incorrectly interpret behaviours such as looking away from a screen, temporary internet interruptions, or environmental distractions as suspicious, potentially resulting in false accusations.

Remote proctoring systems should minimize data collection, provide transparent explanations of monitoring practices, and ensure that disciplinary decisions are never based solely on automated outputs. Human review should remain mandatory whenever AI systems identify potential misconduct.

5.1.1.4 Incident 158

The rapid expansion of online learning and remote assessment has led many educational institutions to adopt artificial intelligence-powered facial recognition systems to verify student identities before and during examinations. These technologies are designed to minimise academic dishonesty by ensuring that the individual taking the examination is the registered student. AI-powered identity verification systems typically compare a student's live webcam image against previously submitted identification documents or institutional records before granting access to an examination.

However, AIID Incident 158 documents a case in which a facial recognition system reportedly failed to verify a black student's identity during an online assessment. Despite the student following the required verification procedure, the AI system repeatedly failed to recognise the individual's face, preventing timely access to the examination. Rather than improving the assessment process, the technology became an obstacle that delayed or denied participation. Although the system was introduced to strengthen examination security, the incident highlighted that technical failures can themselves create unfair educational outcomes.

This incident illustrates that AI systems should not become single points of failure within educational processes. Institutions should maintain accessible alternative verification methods, such as manual identity verification by trained staff, whenever automated systems fail. This ensures that technology enhances educational processes without compromising fairness or equal access.

The case further reinforces the importance of incorporating AI ethics into AI education. Future AI practitioners should be trained not only to optimize model accuracy but also to evaluate fairness, accessibility, transparency, and accountability throughout the AI development lifecycle.

5.1.1.5 Incident 466

Following the widespread adoption of generative AI tools such as ChatGPT in late 2022, educational institutions began searching for reliable methods to detect AI-generated assignments. In response, several AI text detection systems, including OpenAI's AI Text Classifier, GPTZero, and Originality.ai, were developed to help educators identify whether submitted work had been written by artificial intelligence.

However, public testing and independent evaluations quickly demonstrated significant limitations in these detection tools. According to the AI Incident Database, many AI-generated text detectors produced both false positives, incorrectly identifying human-written work as AI-generated, and false negatives, failing to recognise AI-generated content. OpenAI itself acknowledged that its own classifier correctly identified only a small proportion of AI-generated text while also incorrectly labelling some human-written work as AI-generated.

The primary concern is reliability. Educational assessment depends upon accurate evaluation because grades directly affect students' academic progression, scholarships, university admissions, and future employment opportunities. An AI system that incorrectly identifies authentic student work as AI-generated may unfairly accuse students of academic misconduct despite no wrongdoing.

The incident also raises concerns regarding procedural fairness. Students often have limited opportunities to understand or challenge how AI detection systems reach their conclusions. Many commercial detectors operate as proprietary "black-box" systems, meaning their methods are not publicly explained or independently verifiable (Balasubramaniam et al., 2023). Consequently, students may face disciplinary investigations without access to a transparent explanation of the evidence used against them.

The incident also demonstrates that assessment practices should not become overly dependent upon automated decision-making. AI detection systems may assist educators by identifying submissions requiring further review, but they should never function as the sole basis for determining academic misconduct.

Collectively, these incidents demonstrate that ethical risks in educational AI extend far beyond academic dishonesty. AI will continue to play a significant role in both education and our society. However, providing unscrupulous students the ability to use AI to cheat their way out of struggling with an assignment, learning how to deal with failure, and working hard to get a good grade, without much capacity for the educator to detect them, combined with tools like AI humanisers, etc. is a “recipe for - well, not disaster, but the further degradation of a type of assignment that has been around for centuries” (Peritz, 2022).

5.1.2 AIID Praise in the Literature

A study was conducted by Feffer et al.(2023), on students with a background in machine learning and STEM. Before the activity was conducted, students' priority was fairness and safety in order to make sure that there is responsible use of the algorithm, which was seen through a pre-class questionnaire. Individually, students were asked to explore ten incidents on AIID and then choose one incident that really stood out to them. They then answered a questionnaire on the incident, after which they formed teams of four to look for a recent (i.e., 2019–2023) AI/ML incident story that was not documented in the database. They then submitted an incident report on the undocumented story to the database. Students were then asked to complete a post-activity questionnaire, designed to evaluate the effectiveness of the class exercise. The questionnaire prompted students to share their feedback on the tool and asked them to revisit the questions from the pre-class questionnaire. In particular, the questions were based on whether working with the AIID impacted their response to the questions of the pre-class questionnaire, which were about:

- the topics they wanted to learn about and their motivation to take this course,
- their concerns / excitement for applications of ML is society,
- their prioritization of values,
- and the role of ML experts in forwarding those values

After the activity, students' focus shifted more toward safety, accountability, and governance, moving away from topics like fairness, privacy, and transparency. Many students mentioned that using the AI Incident Database (AIID) was valuable for learning about various incidents, including some that were surprising or even shocking. These examples underscored the importance of prioritizing safety in AI deployment and gave students a stronger sense of the magnitude and seriousness of AI-related issues.

In the post-activity discussion, students proposed a range of solutions and agreed that while steps needed to be taken to prevent incidents from recurring, there were multiple complexities involved. Some suggested using risk-management tools like "datasheets for datasets," while others favored more proactive regulatory approaches, such as human-value alignment or including human oversight in the feedback loop. They also noted that challenges like the lack of explainability and the rapid pace of AI development make tackling these issues hard, but that "thinking carefully about desired outcomes" might help guide progress.

In their open-ended responses, students emphasized that promoting the AIID more widely could encourage broader contributions and usage. They also suggested that speeding up the review process through some automated integration with social media could help AIID users. Additionally, they felt that the AIID would benefit from placing a stronger focus on accountability and expanding the definition of "incident" to include potential risks, not just realized ones, echoing a proposal already under consideration by the AIID team.

AIID is a useful resource for raising awareness of AI harms to students, but there is room for improvement. Firstly, since our students could submit news reports sourced from their social media, one

area for improvement is developing tools that seamlessly integrate reporting with social media feeds. For example, we could create a browser plugin that allows users to report incidents with just a click. Another avenue of improvement concerns the speed and efficiency of the database. Thirdly, most of the incidents are concentrated amongst the US, which shows lack of awareness about this database in other countries. This paper is an attempt to spread awareness about the database, so that the maximum number of countries can be covered and their incidents recorded. Lastly, exploring alternative schemes to leaderboards to incentivize more widespread reporting could lead to ultimately enhancing the database's usage and long-term value. (Feffer et al., 2023).

5.2 Implementing Incident-Based Ethics Education

This paper proposes incorporating AIID case studies into existing AI and computer science curricula rather than introducing an entirely separate ethics course. Incident-based learning can be integrated through classroom discussions, case-study analyses, group presentations, and reflective writing assignments.

A possible four-stage classroom activity is outlined below.

Stage 1: Students individually investigate an AIID incident and identify the stakeholders, AI technology involved, and documented harms.

Stage 2: Working in small groups, students analyse the incident using ethical values such as fairness, accountability, transparency, privacy, and human oversight.

Stage 3: Students propose technical and policy interventions to reduce the likelihood of similar harms.

Stage 4: Students reflect on how the incident influences their own approach to designing responsible AI systems.

By repeatedly analysing documented incidents, students develop the ability to anticipate potential failures before deploying AI systems, helping reduce what Boyarskaya et al. (2020) describe as "failures of imagination."

5.3 Final Recommendations

The findings presented in the previous sections demonstrate the need for AI ethics to become a core component of AI curricula, particularly for current and future AI practitioners developing systems for educational settings. Nowadays, AI education prioritizes technical skills such as programming, machine learning and model performance while ethical considerations take a back seat. However, as seen from this paper, the growing influence of AI on student learning, assessment, feedback and educational decision making makes ethical awareness just as important as technical prowess.

Future practitioners need to be trained to critically evaluate societal and educational consequences of their technologies, and not just focus on building useful and efficient AI systems. Without awareness or insight into how their system could fail, developers may unintentionally deploy systems that practice bias, reinforce discrimination, compromise student growth, or create harmful educational environments. A recurring theme throughout this paper is the problem of *failures of imagination*, in which developers fail to anticipate how AI systems may cause harm once deployed in real-world settings (Boyarskaya et al., 2020). This is why past incidents where AI has caused harm need to be studied. This is where tools such as the Artificial Intelligence Incident Database (AIID) provide students and practitioners with real-world examples of algorithmic failures, helping them better recognize potential risks before deployment. Introducing incident-based ethics education early in AI curricula may help reduce failures of imagination by exposing students to documented examples of AI harms before they design and deploy AI systems. However, this educational approach requires further empirical evaluation. Jakesch et al. have found that AI practitioners prioritize fairness over safety whereas Laufer et al. report that the AI ethics community prioritize accountability over fairness. After exposure to the AIID, these views may change as students and creators understand the consequences of algorithms and their potential harms (Feffer et al., 2023). Ultimately, ethical reflection cannot be treated as an optional addition to AI education, but as a foundational component necessary for the responsible development of AI technologies in education and beyond.

6 LIMITATIONS

This paper has several limitations. First, the discussion of AI incidents is based on a relatively small number of representative cases rather than a comprehensive analysis of all incidents contained within the Artificial Intelligence Incident Database (AIID). Although these examples illustrate important ethical issues, they should not be interpreted as exhaustive of the challenges associated with AI in education.

Second, the paper relies primarily on AIID incident reports and secondary sources rather than original empirical data, such as interviews with educators, classroom observations, or surveys of students. Consequently, the analysis reflects documented reports and published literature rather than first-hand evidence from educational settings.

Third, the AIID itself has limitations. As noted by Feffer et al. (2023), many reported incidents are concentrated in the United States, suggesting that the database may underrepresent harms occurring in other countries because of differences in reporting practices and public awareness. The database should therefore be viewed as an important but incomplete record of AI-related harms.

Finally, this paper proposes incorporating AIID case studies into AI ethics education, but it does not empirically evaluate the effectiveness of this curriculum. Future work should implement and assess classroom interventions to determine whether exposure to real-world incidents improves students' ethical reasoning, awareness of AI risks, and decision-making.

7 CONCLUSION

This paper argues that AI raises ethical concerns when applied to education. Primarily these concerns were AI-generated feedback and compromised student growth. The paper further proposes mutltiple solutions including the AIID. The AIID is a useful tool in providing solutions to raise awareness about incidents where AI has caused or nearly caused harm and to prevent further such incidents from repeating. AI is a new tool; its impact depends on the choices humans make when designing, deploying, and governing these systems. Furthermore, it is only going to play an increasing role in our lives given the pace of discovery of new use cases for AI.

When AI causes harm, humans need to step up and take responsibility. AI ethics should become a foundational component of AI education rather than an optional addition to technical curricula. Teaching future AI practitioners solely how to develop increasingly capable systems is no longer sufficient; they must also understand the ethical responsibilities that accompany those capabilities. Incorporating incident-based ethics education into AI curricula allows students to analyse real-world failures, critically evaluate the decisions that contributed to those outcomes, and develop the skills necessary to anticipate and mitigate similar risks in future AI systems. Such an approach aligns closely with the precautionary principle by encouraging developers to identify and address potential harms before they affect users.

At the same time, AI should not be viewed as a replacement for educators but as a tool that supports human expertise. Human-AI collaborative models, in which AI assists with teaching, assessment, and administrative tasks while educators retain meaningful oversight and professional judgement, offer the most responsible path forward.

8 ACKNOWLEDGEMENTS

I would like to thank my mentors, Samar Sabie and Lorenzo Elijah, for all their guidance, support and help during the writing of this paper.

REFERENCES

1. Feffer, Michael, Nikolas Martelaro, and Hoda Heidari. 2023. *The AI Incident Database as an Educational Tool to Raise Awareness of AI Harms: A Classroom Exploration of Efficacy, Limitations, & Future Improvements*. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*. ACM. https://doi.org/10.1145/3617694.3623223
2. Jakesch, Maurice, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. *How Different Groups Prioritize Ethical Values for Responsible AI.* In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM. https://doi.org/10.1145/3531146.3533097
3. Laufer, Benjamin, Sameer Jain, A. Feder Cooper, Jon Kleinberg, and Hoda Heidari. 2022. *Four Years of FAccT: A Reflexive, Mixed-Methods Analysis of Research Contributions, Shortcomings,*

and Future Prospects. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22).* ACM. https://doi.org/10.1145/3531146.3533107

4. Peritz, Aki. 2022. *A Fun, Easy New Way for Students to Cheat. Slate*, September 6, 2022. https://slate.com/technology/2022/09/ai-students-writing-cheating-sudowrite.html
5. Artificial Intelligence Incident Database. 2022. *Incident 339: Open-Source Generative Models Abused by Students to Cheat on Assignments and Exams.* https://incidentdatabase.ai/cite/339/
6. Artificial Intelligence Incident Database. 2013. *Incident 135: UT Austin Grade Algorithm Allegedly Reinforced Historical Inequalities.* https://incidentdatabase.ai/cite/135/
7. Artificial Intelligence Incident Database. n.d. *The Artificial Intelligence Incident Database.* https://incidentdatabase.ai/
8. Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2019. *Improving Fairness in Machine Learning Systems.* In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM. https://doi.org/10.1145/3290605.3300830
9. Boyarskaya, Margarita, Alexandra Olteanu, and Kate Crawford. 2020. *Overcoming Failures of Imagination in AI-Infused System Development and Deployment.* arXiv:2011.13416. https://arxiv.org/abs/2011.13416
10. Lynch, Lori. 2024. *AI in Education: Introduction. Britannica Education.* https://britannicaeducation.com/blog/ai-in-education/
11. Idowu, Jamiu Adekunle. 2024. *Debiasing Education Algorithms. International Journal of Artificial Intelligence in Education.* https://doi.org/10.1007/s40593-023-00389-4
12. *The Four Pillars of Great Assessment.* 2021. *The Learning Review,* St George's British International School. https://issuu.com/stgeorgesrome/docs/the_learning_review_-_spring_2021/s/11958253
13. Baker, Ryan S., and Aaron Hawn. 2021. *Algorithmic Bias in Education. International Journal of Artificial Intelligence in Education.* https://doi.org/10.1007/s40593-021-00285-9
14. Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. *Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application.* https://doi.org/10.1146/annurev-statistics-042720-125902
15. Gardner, Josh, Christopher Brooks, and Ryan Baker. 2019. *Evaluating the Fairness of Predictive Student Models Through Slicing Analysis.* In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge.* ACM. https://doi.org/10.1145/3303772.3303791
16. Suresh, Harini, and John Guttag. 2021. *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle.* In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21).* ACM. https://doi.org/10.1145/3465416.3483305
17. Balasubramaniam, Nagadivya, Marjo Kauppinen, Antti Rannisto, Kari Hiekkanen, and Sari Kujala. 2023. *Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements. Information and Software Technology* 162. https://doi.org/10.1016/j.infsof.2023.107197
18. Sharma, M., M. Tong, T. Korbak, D. Duvenaud, A. Jermyn, S. Cheng, et al. 2023. *Towards Understanding Sycophancy in Language Models.* arXiv:2310.13548.

19. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT.* arXiv. https://doi.org/10.48550/arXiv.2302.11382
20. UNESCO. (2005). *The Precautionary Principle.* World Commission on the Ethics of Scientific Knowledge and Technology (COMEST). Paris: UNESCO.
21. Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). *Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements.* Information and Software Technology, 159, 107197. https://doi.org/10.1016/j.infsof.2023.107197