An Ethical Impact Assessment for the Use of Emotion Recognition AI in Education

Daniel Mason

***Index Terms*—AI, Ethics, Bias, Emotion recognition, Value sensitive design**

I. Introduction

THIS company is grounded in the use of emotional recognition for monitoring student engagement and understanding within an academic space. By identifying basic emotions general well-being can be monitored. This algorithm, however, will be designed to recognise emotional indicators tailored for educational success. Recognising these indicators could alert educators to adjust their teaching methods to introduce interactive activities to re-engage students. They may also help educators tailor lessons to capitalise on students' curiosities, deepening their learning as well as informing educators about the effectiveness of their teaching methods.

In the realm of emotion recognition, it is important to concurrently consider the development of the technology and the understanding of the ethical implications associated [1]. Outlining an ethical impact assessment (EIA) ensures adherence to legal regulations as well as aligning to societal values, mitigating potential harm to users and ensuring the safety of the company. A method that systematically includes ethical considerations is value sensitive design (VSD). This is an approach towards technological progress that emphasises the integration of human values into the development of the given technology by employing a tripartite method where conceptual, empirical and technical investigations are undertaken. [2] The initial step in utilising VSD is identifying priority values of the project and stakeholders, this can be complicated as some values can conflict. Therefore, the ability to compromise and prioritise becomes essential in crafting an optimal design that effectively balances benefits and trade-offs. By utilising this approach a company can minimise the risk to the user and build itself as a reputable enterprise that aligns with a strong ethical engagement.

In the context of this company, the core values to consider revolve around privacy, consent, fairness and the accuracy of the algorithm. There is also the need for explainability of the Machine learning systems.

II. Ethical impact assessment informed by Value Sensitive Design

TABLE I: Definitions of human values [3]

Value	Definition
Privacy	The right to choose what information about oneself is communicated to others
Efficacy	Ability to produce the desired result
Consent	Refers to gaining people's agreement and a disclosure of understanding where voluntariness and competence are necessary
Fairness	Freedom from a systematic bias against individuals or groups, in this setting fairness would be that variation in performance across diverse groups is as small as possible
Accountability	Ensures that the actions of an individual or organization may be traced
Explainability	The ability to explain an ML model in human terms so that a model is better understood
Autonomy	The right to self-decision and action
Trust	An expectation of goodwill between parties
Transparency	Making access to all relevant information accessible and justifications for decisions and actions

To help understand the priorities of the stakeholders, an analysis of their respective needs was undertaken and is detailed in TABLE II and the following text, this was done in conjunction with a literature review of other papers to review published interpretations of stakeholders needs where commonly associated values were taken into consideration.

For teachers the efficacy of the product alongside the well-being of the students are the priority values, this is because the teachers' aim is to help their pupils perform without detriment to their mental health and without suppressing their ability to express themselves. For the students themselves, the concerns are their privacy [4] and the fairness of the algorithm. There is a concern for potential prejudices in the algorithm that may falsely or unfairly assess a student's emotional state. Parents as the primary caretaker of the student, are most interested in the transparency of the results and the explainability of

TABLE II: An analysis of the key stakeholders in the project

Stakeholder	Involvement	Priority Values	Motivation	Potential Risks
Educator, primary user of the technology	Direct	Efficacy, Student well-being	Ensuring effective teaching as well as creating a nurturing environment with positive mental and emotional health	Misuse of emotional data or improper interpretations could cause unwanted effects on learning and damage teacher reputation.
Students, subject to the model	Direct	Consent, Privacy, Autonomy, Academic Success, Fairness	Control over their personal information and reserving the right to independent choices regarding their education.	Surveillance may cause an inhibition to authentic expression as well as general discomfort. Vulnerabilities in data security could cause the breach of sensitive data. Algorithm biases may perpetuate unfair judgments
Parents, Responsible for well-being of children	Direct	Student well-being, Academic success, Transparency, Explainability	Wanting access to their childrens' emotional and academic welfare	Transparency concerns, especially if they feel uninformed on how the data is being used
Tech devs, responsible for the design and implementation of the technology	Direct	Ethical design, Fairness, Accountability, Transparency, Trust	Progressive tech with equitable treatment of people	Algorithm biases may perpetuate social inequalities or limit the efficacy of the product
Policy makers	Indirect	Policy alignment, Fairness	Ensuring equitable opportunities and policy alignment	Failures to meet policies could result in legal consequences or cause a loos in trust of the educational system. New policies may be needed to address privacy concerns
Educational auditors	Indirect	Educational efficacy, Ethical use, Privacy, Accountability	Ensures ethical use of the data gathered and that educational practices are effective	Failure in efficacy could mean negative consequences for students

the ML process, as FE news reports, 'a lack of official guidance is leaving parents in the dark' [5]. They want to be able to trust the educational system to be a safe space for their children to learn and as such would be curious to the workings and security of the product. The policy makers will be most concerned with the fairness and privacy such that all students are allowed the same opportunities and treatment. It is expected that AI will have a transformative change on the education system [6] meaning policies are likely subject to change in future. Similarly, educational auditors may focus on the accountability and insurance of the ethical use of the technology. Fortunately, the use of AI in education is shown to hold mostly positive views. [6] It is the tech developer's job to manage these values. As stated before, this will require compromise because of conflicting values. For example, privacy may be a priority for students, but there is an expected transparency from parents. This will have to be managed by the educator as a middle man trusted to raise serious concerns to parents. Ultimately the privacy of the student is prioritised until serious adverse circumstances are apparent. Results of the model are trusted solely to the given educational space and are expected to be used with the students' best interests where parents are to be contacted at the schools trusted judgment.

A. *Future possibilities and concerns*

First-hand empirical investigations such as surveys of stakeholders may provide insight into the expectations and concerns about the use of this AI specifically. Small scale pilot studies could be used to observe the potential impacts, positive and negative that the product may have. The main areas of concern for the project lie in the potential for bias, there is potential for the misinterpretation of emotion depending on a student's background and preexisting biases in society like those seen in the COMPAS algorithm [7]

III. Recommendations and Considerations

A. *Motivation for recommendations*

The EIA demands a system with an understanding and recognition of a diverse array of emotional indicators. As a result data gathering on a widespread scale over a range of emotions and backgrounds will be vital. There is further the consideration that emotional cues may be subtle such as the differentiation between frustration and boredom. To analyse this video captures of classroom settings that can be annotated and learned from will be needed. The EIA highlights privacy and fairness as important values to the stakeholders, to address this further information on the legal frameworks of privacy regarding students will be needed and as expected bias mitigation will be paramount.

B. Datasets

The datasets utilised must represent the diverse student population across a range of categories such as age, ethnicity and background to give the best possibility of accuracy and fairness. The data source used should be acquired from a consenting source where the privacy of all participants is respected. Further, the training data must be clear and accurately labelled in order to produce the most effective model. A great deal of datasets used for emotion recognition are accessible such as the EMOTIC dataset provided by Kosti *et al* [8] this specific dataset contains over 23,000 labelled images of a diverse range of people showcasing 26 discrete categories of emotions. Despite these benefits this dataset would prove unsuitable for this task as it is used for the adult population and it discloses that some images are taken from Google images, as such it is unknown whether the participants are consenting for their images use in our model's training. Therefore it is likely that primary collection and documentation of data will be necessary. This could be carried out in consenting control groups of students exposed to a range of scenarios aimed to cause a certain emotional reaction which could be later analysed and labelled by experts with an emphasis placed on consistency.

C. Risk and Bias Mitigation

To mitigate the bias of the system we audit the algorithm using fairness assessment toolkits such as IBM's AI fairness 360 [9] which works by producing metrics that evaluate the fairness visually and help understand the behaviour of the model across a range of categories e.g. gender, age and race meaning any systematic biases are identified and can be dealt with, actively addressing concerns related to fairness.

Once identified, adjustments to the algorithm can be made by adding constraints such as a term that penalises the model for errors in the identified biased groups. We could further help diversify and aid the model by periodically making the training dataset larger even after it is already being used. Essentially providing it continuous learning.

Another option could be to train the model with human feedback, where an expert could review the primary predictions it makes and give input that would be used to improve its next iteration.

As identified consent and privacy are vital, to ensure a trusting relationship with all stakeholders, to settle any risks of privacy, data collected will be only what is necessary to facilitate the model. Data taken will be pseudonymised so that information cannot be linked to any student specifically, without knowing the corresponding names which only the educator will need access to.

D. Critical Assessment and Limitations

The data collection process suggested is adequately suited to the needs of the task but raises ethical concerns regarding the use of images of children. As a result it would require consistent agreements of consent from parents and an implicit trust in the teachers and technology. As the data collection is carried out first hand it is limited by scale but all other considerations are within the company's control as such we can adhere to all necessary ethical practices. The data evaluation tool used (IBM AI Fairness 360) is reputable for its use in identifying biases in models and should only be limited by the data provided to it. It is important to recognise that the introduction of new data as suggested may bring with it the introduction of new biases so continued monitoring of the model's performance remains necessary.

Some issues remain difficult to address and need further research, this includes the unknown psychological detriment continuous monitoring could have and the impact on a student's freedom of expression. Overall, while the suggestion has promise in theoretical use, it comes with practical limitations and would require incredible care to remain ethically sound if pursued.

References

[1] Unesco, "Ethical impact assessment: A tool of the recommendation on the ethics of artificial intelligence," 2023.

[2] P. Batya Friedman, H. Kahn, and A. Borning, "Value sensitive design and information systems," *The Ethics of Information Technologies*, 2020.

[3] P. Zhang and D. Galletta, *Human-computer Interaction and Management Information Systems: Foundations: Foundations*. Taylor & Francis, 2015.

[4] S. Akgun and C. Greenhow, "Artificial intelligence in education: Addressing ethical challenges in k-12 settings," *AI and Ethics*, vol. 2, p. 431–440, Sep 2021.

[5] F. N. Editor, "Schools and parents unprepared for ai revolution," Feb 2024.

[6] J. Felix, "Use of artificial intelligence in education delivery and ...," 2024.

[7] A. Meshi, “Deconstructing whiteness: Visualizing racial bias in a face recognition algorithm,” *10th International Conference on Digital and Interactive Arts*, Oct 2021.

[8] R. Kosti, J. Alvarez, A. Recasens, and A. Lapedriza, “Context based emotion recognition using emotic dataset,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2019.

[9] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, “Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias,” 2018.