Lee, A. V. Y., Luco, A. C., & Tan, S. C. (2023). A Human-Centric Automated Essay Scoring and Feedback System for the Development of Ethical Reasoning. *Educational Technology & Society*, *26*(1), 147-159. https://doi.org/10.30191/ETS.202301_26(1).0011

A Human-Centric Automated Essay Scoring and Feedback System for the Development of Ethical Reasoning

Alwyn Vwen Yen Lee[1*], Andrés Carlos Luco[2] and Seng Chee Tan[1]
[1]National Institute of Education, Nanyang Technological University, Singapore // [2]Nanyang Technological University, Singapore // alwyn.lee@nie.edu.sg // acluco@ntu.edu.sg // sengchee.tan@nie.edu.sg
[*]Corresponding author

ABSTRACT: Although artificial Intelligence (AI) is prevalent and impacts facets of daily life, there is limited research on responsible and humanistic design, implementation, and evaluation of AI, especially in the field of education. Afterall, learning is inherently a social endeavor involving human interactions, rendering the need for AI designs to be approached from a humanistic perspective, or human-centered AI (HAI). This study focuses on the use of essays as a principal means for assessing learning outcomes, through students' writing in subjects that require arguments and justifications, such as ethics and moral reasoning. We considered AI with a human and student-centric design for formative assessment, using an automated essay scoring (AES) and feedback system to address issues of running an online course with large enrolment and to provide efficient feedback to students with substantial time savings for the instructor. The development of the AES system occurred over four phases as part of an iterative design cycle. A mixed-method approach was used, allowing instructors to qualitatively code subsets of data for training a machine learning model based on the Random Forest algorithm. This model was subsequently used to automatically score more essays at scale. Findings show substantial agreement on inter-rater reliability before model training was conducted with acceptable training accuracy. The AES system's performance was slightly less accurate than human raters but is improvable over multiple iterations of the iterative design cycle. This system has allowed instructors to provide formative feedback, which was not possible in previous runs of the course.

Keywords: Automated essay grading, Human-centric AI, Formative feedback, Machine learning, Ethics education

1. Introduction

Over the past decades, the deployment of Artificial Intelligence (AI) has transited from a nascent idea into an established field that is widespread and undeniably impactful on education with profound possibilities (Holmes et al., 2019). With untapped potential to create impacts by augmenting human intelligence with machine intelligence for educational research and purposes (Yang, 2021), there is also growing research on how AI can sustainably do so (Vinuesa et al., 2020). However, even though the advancement of AI entails the need to enable human welfare by improving human conditions, there remains a critical need to investigate how AI can be responsibly designed, implemented, and evaluated, especially in the field of education. Afterall, learning is inherently a social endeavor involving human interactions and not just disembodied human-machine transactions (D'Mello, 2021), rendering the need for AI designs to be approached from a humanistic perspective as human-centered AI (HAI) with consideration of human conditions and contexts (Yang et al., 2021).

This is more recently viewed to be an emergent and urgent concern, as an increasing number of functions in AI systems have already been ceded to algorithms to the detriment of human control, resulting in growing unease and loss of equitability (Sareen et al., 2020). Further, as a new-age workforce constantly evolves with a constant flux of expectations and needs, the identification of potential knowledge gaps and deficits of expertise in higher education can help support the development and implementation of AI in education (Lee, 2020). Students can remain relevant in the new reality by equipping themselves with literacies and skills to thrive in new economies while teachers adapt to new models and orientations to accommodate lifelong learning (Aoun, 2017). It is therefore unsurprising to note that a growing number of recent studies and meta-studies have utilized AI-supported systems (e.g., Garcia-Magarino et al., 2019; Lepri et al., 2021) but are focusing more on trustworthy systems with explainable layers, so that users have the opportunity to understand the reasons behind decisions. AI designs should also then consider human conditions and have a human-oriented approach when augmenting human intelligence with machine intelligence (Yang et al., 2021).

Students as future leaders will face the above-mentioned challenges as AI continues to shape society. Therefore, considering how students navigate the existent knowledge society, the study of ethical reasoning plays a key role in enhancing students' problem-solving capacity and exercising their minds in the disciplines of critical and logical thought. However, the use of AI in the domain of philosophy remains limited due to differences in

philosophical, pedagogical, and technological approaches. On one hand, it may be surprising to some that most AI work does not require any philosophy since a restricted representation has already been designed or programmed (McCarthy, 1995). On the other hand, this should not detract from the potential of using AI in studies of philosophy, of which the ease of study can allow both teachers and students in higher education to better adapt to new ways of teaching and learning. Even so, emergent societal needs such as sustainable assessment for lifelong learning will require significant shifts away from the current focus of assessment of learning (summative assessment) to assessment for learning (formative assessment) (Nguyen & Walker, 2016).

When undergoing this transition, the successful use of AI in the form of Automated Essay Scoring (AES) within the field of summative assessment of learning (Gardner et al., 2021) can offer exciting opportunities for formative assessment. Gardner and his co-authors opined that "AI in educational settings has changed little in its basic percepts and functions" and lamented that machine learning and actions have not delved far beyond intelligent analysis of large-scale data in the last decade. Thille et al. (2014) argued that large-scale assessment should benefit learners by providing continuous, multi-faceted feedback. In this regard, recent advances in AI technologies afford opportunities for formative assessment at scale, such as using machine learning to determine the quality and distribution of ideas in classroom discourse (Lee, 2021) and using trace data to dynamically give young learners immediate performance feedback in comprehension tasks (Walker et al., 2017).

To address these challenges, we attempt to answer how HAI can be designed and used for formative assessment, using processes that adjust algorithms through human contexts and social phenomena. The context of this study represents a situation that is prevalent in many foundational undergraduate courses, which are often offered to large cohorts of students. The course in this study, "Ethics and Moral Reasoning," has an overwhelming number of student registrants, often ranging over 600 students each semester. With these students trekking through an online module that is often supervised by few instructors, several imminent problems became apparent: (1) The course has to be conducted online due to the large number of students, which further enlarges the perceived distance between the instructors and students; (2) for every assignment issued to students to gauge their understanding and progress of learning, the number of returned assignments is overwhelming for a small team of instructors to score accurately and in a timely manner; (3) providing personalized and meaningful feedback to students becomes nearly impossible; and consequently, (4) some students may not be able to grasp the importance and significance of ethics and moral education based on limited interactions with the instructors.

In this study, we use a mixed-method approach consisting of human-designed rubrics for assignment coding, peer assessment and application of machine learning as part of an automated essay scoring and feedback system for the development of ethical reasoning. In response to the challenges of courses with large enrolments that are conducted in an online format, this study attempts to answer the following research question: *To what extent can an automated essay scoring and feedback system be employed to provide formative feedback and potentially act as a surrogate for instructor interactions?*

Addressing this question will benefit the teaching and learning in courses with large enrolments, especially when more online courses are being added to the university's offerings due to the emerging dynamics of the educational landscape and deployment of educational technologies. A caveat is that the deployment of AI, in the context of education practices and computing development, will likely not take over the role of the instructors, due to how teaching and learning happen in the classroom and the ways in which it is profoundly different from human intelligence that AI seeks to emulate (Cope et al., 2020). More importantly, tools and systems modified through this study can focus on learning from human inputs and collaborations, to support course designs that focus on improving humanistic aspects such as students' communications and critical thinking skill development, through the provision of formative feedback that is more timely, meaningful, and actionable.

2. Background

In this section, we highlight the importance of education in ethics and moral reasoning, followed by the significance of formative feedback, an overview of several automated essay scoring and feedback systems, and lastly, our selection of method in this study.

2.1. Importance of education in ethics and moral reasoning

For an emergent knowledge society to assimilate meaningful use of AI for teaching and learning, students will need to develop ethical reasoning to enhance problem-solving capacity and exercise minds in the disciplines of

critical and logical thinking. In an ideal situation, courses in ethics put students on paths toward what Lawrence Kohlberg, a famous psychologist, termed "postconventional" moral reasoning (Rest et al., 1999). At this stage of moral reasoning, "individual judgments are now determined by self-chosen, internal principles rather than accepted from external authorities" (Vozzola, 2014, p. 29). To cultivate skills in postconventional moral reasoning, students should have ample opportunities to express their values. More importantly, they should be challenged to defend and refine their values in response to feedback from others. By participating in a university course in ethics, students are not just introduced to moral values that one or another thinker believes in, they are also challenged to reflectively cultivate their own values. They are given sufficient space and opportunities to express themselves and defend or refine their values and opinions in response to others.

In addition, as McKeachie and Svinicki (2006) noted, "values are not likely to be changed much simply by passively listening and observing a lecturer. Change is more likely in situations in which the teacher and the students reflect, listen, and learn from one another" (p. 338). In order to develop good values and live reliably by them, one needs to develop skills in moral reasoning, which is the ability to independently assess a situation, identify morally relevant considerations, and arrive at judgments about what one ought to do. Thus, in wanting to be ethical during undergraduate studies and in the society that awaits students after graduation, they have to be able to think through complex moral situations by themselves and rely on their own powers of moral reasoning. The course that students undertake in this study is an opportunity and setting that provides a sampling of such situations. For such a course, regular formative assessment and feedback provided by peers and instructors are deemed to be critical.

2.2. Significance of formative feedback for teaching and learning

Formative feedback is defined as the "information communicated to the learner that is intended to modify his or her thinking or behavior to improve learning" (Shute, 2008, p. 153) or in layman terms, is any message delivered to a learner while there is still time to adjust. Receiving feedback challenges learners' existing beliefs and forces them to evaluate their positions. Formative feedback is not limited by fields and can be relevant in the sciences (Shavelson et al., 2008), engineering (Roselli & Brophy, 2006), the humanities and life in general (Shute, 2008). In general, formative feedback can be provided in many ways, from teachers' written feedback to full critique sessions of an engineer's work-in-progress (Shute, 2008).

A meta-analysis conducted by Hattie (2013) found that among all the pedagogical methods that instructors have at their disposal, the provision of formative feedback consistently yields one of the most powerful effect sizes. Formative feedback, when used for the enhancement of learning and achievement, can help instructors realign their teaching in response to learners' needs (Juwah et al., 2004). The importance of formative feedback cannot be overstated as it motivates learners to take greater agency in their learning, and potentially provides a direction for improvement. Although feedback is among the major influences, the type of feedback and the way it is given can be differentially effective (Hattie & Timperley, 2007), such as the timing of feedback and both positive and negative impacts on learning.

When providing effective formative feedback for teaching and learning purposes, essay writing is considered one of the most useful tools for either assessing students' learning, their ability in organizing and integrating of ideas into a knowledge artifact, or the competency of expressing oneself in writing (Valenti et al., 2003). The scoring of free-written responses such as essays, however, is a non-trivial process with inherent challenges such as the perceived subjectivity of the grading process. Hence, this problem attracted a large range of methods and techniques as solutions, including neural approaches (e.g., Taghipour & Ng, 2016), techniques such as Bayes' theorem (e.g., Rudner & Liang, 2002), and more prevalently natural language processes involving semantic analysis (e.g., Rokade et al., 2018) to grade free form texts or essays.

From the instructors' perspective, the availability of technology does alleviate parts of the teaching load, but there remains potential pedagogical impediment that affects instruction and assessment. For example, apart from administrative duties, instructors are still expected to handle large groups of students (i.e., lopsided student-teacher ratio) with insufficient scaffolds or tools to facilitate meaningful teaching and learning. To be responsible for the learning needs of a large group of students, it is extremely challenging for instructors to contextualize assignments, correct misconceptions, and still provide timely and accurate feedback – practices that are beneficial for students' personal growth (Hattie & Timperley, 2007). To mitigate the severity of such issues, prior research has suggested peer reviews and evaluations as possible strategies that prompt students to complete assignments in a diligent manner (Liu & Carless, 2006). An alternative to other potential solutions, including expansion of teaching teams or leveraging on peer reviewers, is to automate the grading processes within the

course, at least partially, so that more time and space are freed up for instructors to set up well-established routines that provide feedback to students.

2.3. Automated essay scoring and feedback systems

Automated essay scoring (AES) systems attempt to accomplish part of what instructors do in assessment – evaluate students' work and provide feedback to the students. Even as far back as several decades ago, the goal of these systems has always been to make them at least indistinguishable from human raters (Page, 1966). The eventual goal of these systems is to deliver a consistent assessment comparable with human graders. To develop an AES system, a large dataset is often split up into smaller subsets of data, with some subsets allocated for training and the remainder for validation and testing. The system firstly utilizes the marks scored by experts (which in many cases are the instructors) as labels, then attempts to generate models based on the source material (essays), before using the models to score the remaining essays in the dataset. To approach expert levels of analysis and accuracy, additional training sets labelled with expert ratings are often used in multiple passes of the training dataset, also known as epochs.

The field has developed much since Ellis Page and his colleagues developed the first AES system, Project Essay Grade (PEG), for college-level and adult writers (Page, 1966). Essentially, like many current AES systems, measures are used to approximate features of interest and describe the quality of essays designed for students and writing practice. Since then, several prominent commercial AES systems have been developed and improved, such as E-rater (Attali & Burstein, 2006), Intelligent Essay Assessor (Foltz et al., 1999), and Intellimetric (Elliot, 2003). These AES systems, similar to PEG, assist teachers in the process of essay scoring by allowing students to write and submit their work before the system provides automated feedback. The systems are mostly capable of scoring the essays and providing suggestions for improvement in targeted areas such as language, style, and sentence structure.

For example, Educational Testing Service (ETS) has used E-rater since 1999, based on intuitively small and meaningful features to score essays (Attali & Burstein, 2006). These micro-features produce a single scoring model that can be used across different assessment prompts. It also allows easier modification and upgrading of the system, so as to boost overall grading performance. Common features include grammar, word usage, sentence mechanics, style, lexical complexity, and prompt-specific vocabulary usage. Upgrades to E-rater were designed to flag anomalous and bad-faith essays, which are not scored, while scores for other essays are calculated using a weighted average of standardized feature values followed by a linear transformation to achieve the desired scale. A distinguishing factor E-rater has over other AES is the possible use of judgmental control by end-users, enabling users to determine relative weights, either by content experts or by settings weights based on similar assessments, hence preventing extreme relative weights from affecting the validity of scores.

Another example is the Intelligent Essay Assessor (IEA) (Foltz et al., 1999), which provides an alternative to using an expert training set. IEA is based on Latent Semantic Analysis (LSA) (Landauer et al., 1998). Using domain-representative texts like textbooks, articles, and samples of writing for training, LSA derives a high-dimensional semantic representation of the information within the domain by using vectors, often referring to lists or columns of scalar real numbers, to represent the words and semantic information found in the source material. Vectors may represent sources like student essays and the closer vectors are to each other, the more similar the essays are. Hence, scores for essays can then be determined by comparing each essay against all previously graded essays of similar vector weights. The result is a holistic determination of essay quality and this system can also be used as a generalized solution that extends to subjects such as psychology, biology, and history as well as ETS's Graduate Management Achievement Test (GMAT). Past results have shown that IEA's reliability is comparable to that of human graders, with other features including flagging anomalous essays and essays that are too similar to each other or textbooks, indicating possible instances of plagiarism. However, as much as LSA is used as a formative assessment tool, IEA is not originally designed to be a teaching tool. It compares texts based on semantic similarity, but it cannot assess creative writing or point writers toward improvement of their texts.

The Intellimetric is a proprietary intellectual asset protected by Vantage Learning (Elliot, 2003). The system parses text to flag the syntactic and grammatical structure of essays. The sentences are then tagged for parts of speech, grammatical structures, and concept expressions. Unique words and concept networks are subsequently employed for spelling recognition and grammar checking. The data is coded to form models and these models are then associated with features extracted from text and tentative scores may be assigned. Optimization is eventually performed to provide a single grade to the essay. The robustness in this system comes in the form of

using different models to grade the essays, similar to how multiple judges are employed to conduct manual essay grading.

2.4. Choice of AES system for this study

In summary, the above-mentioned are commercially-ready and established AES systems that provide many benefits to users over multiple iterations of design and improvements, but many of them are proprietary and closed source, or are platform-dependent and require conforming usage to a specific system. The goal of this study is to determine to what extent and form can an AES system exist to combine quantitative features with essay content to provide a reliable method for scoring, and potentially in place of instructors. For social scientists, since developments in algorithms are useful only to the extent that they can access the implementation (Schonlau & Zou, 2020), therefore, machine learning algorithms that are open-source and based on supervised learning models provide an intermediate solution for solving problems that are difficult to solve via conventional programs but are yet able to learn without being explicitly programmed.

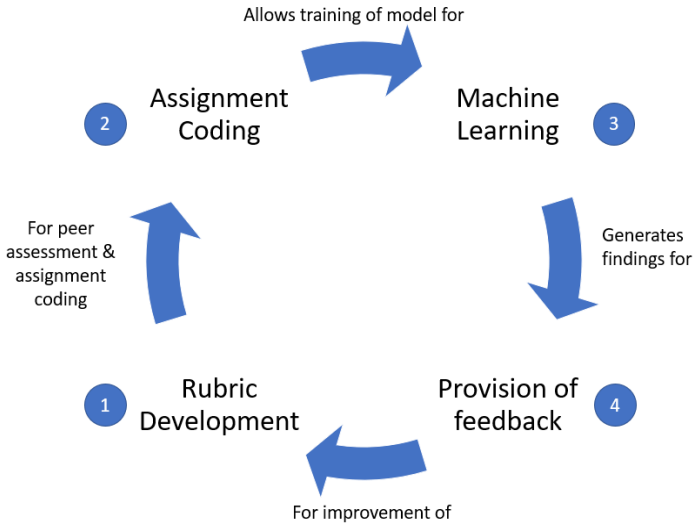
This solution was sought due to the ability to monitor the performance of models and adjust parameters whenever necessary based on the accuracy of prediction. Due to the need for a score-based system, regressions are chosen to be used and among a list of regression algorithms, several studies (Ghanta, 2019; Liu et al., 2012; Schonlau & Zou, 2020) have shown that the random forest models tend to have better prediction accuracy than other regression algorithms (e.g., linear regression, logistic regression, support vector regression) over multiple sets of data. It also fits well with the iterative design cycle (Figure 1) that will be described later in the next section.

3. Method

3.1. Research design

A mixed methods approach was used, involving the analysis and evaluation of qualitative measures during peer assessment and assignment coding, and the use of quantitative methods from the machine learning application. It was part of an iterative design cycle, which is commonly a design-implement-evaluate cycle, where data and analyses in this study can inform and improve the design and scope of learning interventions during subsequent cycles. The provision of a closing feedback loop caters to how we can evaluate the broadening of the study's reach to incorporate other types of learning activities and courses. In this study, we iterated once through the cycle to illustrate the four phases that sequentially occur during the development of the AES system for an actual undergraduate course at a university. These four phases are: (1) Rubric development for peer assessment and assignment coding; (2) assignment coding by instructors; (3) machine learning application; and (4) follow-up actions in providing feedback to students. These phases are further detailed in the following subsections.

Figure 1. The four phases of the iterative design cycle



3.2. Settings and data

In this study, over 600 undergraduate students from across the university attended the course "Ethics and Moral Reasoning," with the entirety of the course being delivered online and split into 13 sessions, also known as units in this study. These units include three major ethical theories' utilitarianism, Kant's deontology, virtue ethics, ethical principles for academic integrity and research, and ethics for sustainability and conservation. Students sequentially progress through the 13 units, at a pace of one unit per week throughout a semester.

The majority of the units began with a short video lecture that covered a well-defined topic in the domain of ethics and moral reasoning, followed by a short selection of readings. As part of students' contributions to the course, each student was requested to either write a short essay (more than 100 words) to a question or to provide a short response (also in short essay format) to another student's essay during some point in the course. Students were encouraged to contribute at least once throughout the course, which led to responses being distributed across the course units. The description and distribution of the essays in the entire dataset is shown in Table 1. These writing assignments also became an entry point for the introduction of student-centered formative feedback.

Table 1. Description and distribution of essays throughout the dataset, with no essays required for the 1st unit

Course unit ID	Number of essays	Average length of each essay (words)
2	781	204
3	193	216
4	184	216
5	52	224
6	117	223
7	357	235
8	531	184
9	99	222
10	159	217
11	108	228

In this study, course unit 2 was selected because it has the highest number of essays. The selected course unit discussed about "Utilitarianism," which referred students to a theory of morality that prescribes actions which maximizes happiness and wellbeing of individuals. For this topic and within the specific week where data collection was conducted, students wrote a total of 781 short essays about utilitarianism or in response to their peers' essays, using the Discussion Board page on the Blackboard learning management system (LMS).

3.3. Rubric development for peer assessment and assignment coding

The team of instructors developed a set of rubrics (see Table 2) with defined criteria to provide guidance to instructors and students during their processes of assignment coding and peer assessment respectively. The rubrics could be used by students to guide their learning from peers and aid self-reflection of own work and were also used by instructors to code and score the short essays, which in turn became the training data for developing the machine learning model. These two processes were independent and did not affect each other: the students obtained formative feedback from peers, while the instructors provided inputs for the machine learning model.

The assignment coding process was conducted by two raters to address the consistency of the proposed implementation and to also obtain a measure of interrater reliability. Due to practical reasons in needing to grade thousands of students with ten units of assignments each, the scoring system was simplified for instructors to use the rubric with the four criteria as a guideline and the theory of majority rule to provide an eventual score of 1 if the essay fulfils the majority of criteria and 0 if otherwise. Essays that received conflicting scores were returned to the pair of raters for rescoring.

Table 2. Rubrics for peer assessment and assignment coding

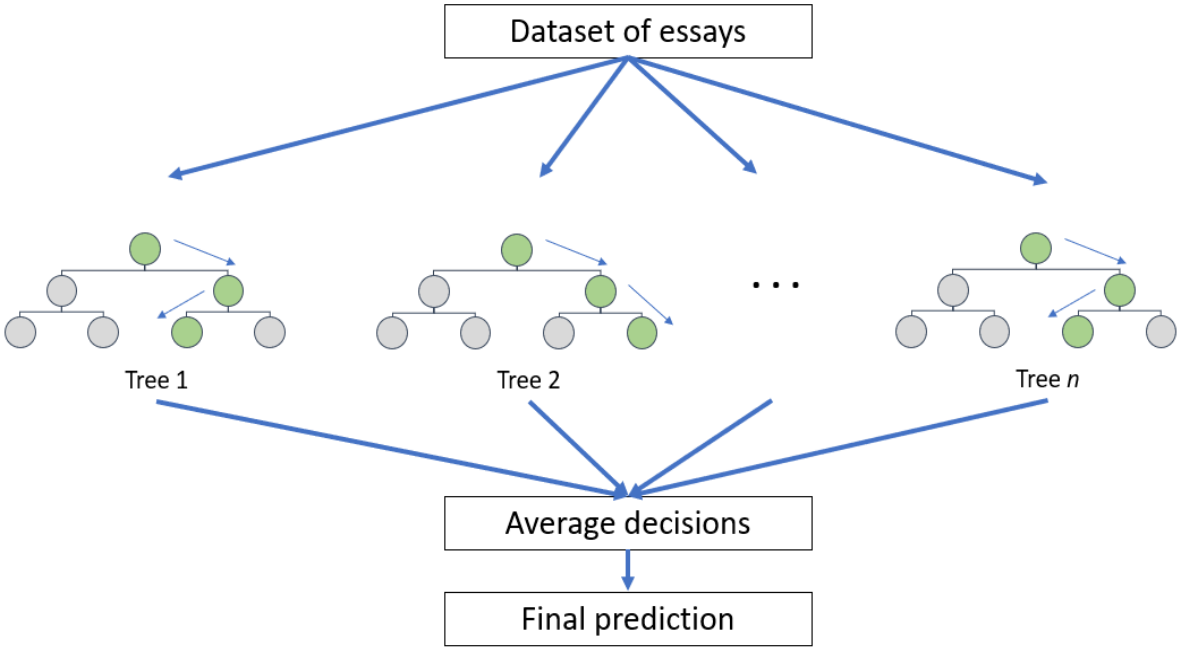
Defined criteria	Criteria grade			
	Excellent	Good	Needs improvement	Poor
Relevance – Content addresses discussion question	Very relevant to the discussion question.	Relevant to the discussion question, contains some digressing content.	Somewhat relevant to the discussion question, with some off-topic content.	Off-topic and no relevance to the discussion question.

Comprehension – Ability to accurately explain key concepts	Accurately explains key concepts necessary for responding to the question, with the use of critical keywords – e.g., utilitarianism; consequences etc.	Largely accurate in explaining key concepts with some minor flaws, with the use of the following keywords – e.g., utilitarianism; consequences etc.	Reflects major misunderstandings of key concepts, failing to refer to relevant key concepts.	Displays no understanding of key concepts.
Thesis – Central statement with at least one argument	Clear and explicit thesis statement.	Explicit thesis but not clearly stated.	Thesis is present but not clear and explicit.	Lacks a thesis statement.
Arguments and reasons for thesis	Clearly lists some pros and cons, weighed against each other to support thesis, preferably with examples.	Pros and cons are suggested but not clearly stated or are not weighed against each other to support the thesis.	Either pros or cons are provided but not both.	Lacks any effort to provide pros and cons.

3.4. Machine learning – Random Forest classifier

The instructors coded a subset of the entire dataset from the course unit "Utilitarianism," which was then used as the training set for an open-source AES system that is modified to work with the LMS in the university. Simply put, if the AES system manages to train a model that has an acceptable level of accuracy (also known as training accuracy) based on a reasonable interrater reliability measure, the model will then be accepted for testing with the remaining parts of the entire dataset before evaluating whether the eventual model can be used for deployment in the LMS.

Figure 2. Essays are processed using multiple trees in a Random Forest algorithm before the decisions are averaged and reach a final prediction



In this study, the Random Forest classifier (Breiman, 2001) was used as a supervised learning algorithm that utilizes the ensemble learning method for regression, by building multiple decision trees and merging them in order to achieve a more accurate and stable prediction. Python was used to code the entire process and the ensemble method of seeding a forest of decision tree learners started with individual decision trees that were grown at random and acted as weak learners, with each tree presenting an outcome that was then coupled together with other trees to create a final forest. When the decisions of this forest were averaged, Random Forest

determined the weight of trees that would be used in the final model for prediction, which could then be utilized for automated scoring of unlabeled essays. This process is visually represented in Figure 2.

However, before the classifier can be implemented, several steps need to be conducted. These include firstly extracting textual data from a multitude of essays on the LMS, before running it through a spelling correcting process such as Norvig's spelling corrector (2007). This is part of preprocessing to avoid interfering with semantics and also because spelling and grammar in this study were not a major consideration in the scoring process. This was followed by feature engineering and extraction to generate multi-dimensional feature vector representations for each essay and outputting into feature arrays, before model training could take place. Random Forest was then implemented to generate a model, after which Cohen's Kappa value (Cohen, 1960) was used as a measure of agreement between the human rater and how well the model predicted using the machine learning algorithm, with correction for chance agreement.

3.5. Provision of feedback

Formative feedback was given to students in two formats. The first kind of feedback was the aforementioned score that was built into the AES system, allowing the instructors to provide a score, albeit a binary one, for every essay that was written by students. This feedback mainly serves to recognize students' effort in thinking, writing, and responding to online discussions, further encouraging them to continue sharing and discussing their ideas and thoughts with fellow peers. This was considered an improvement over the previous iterations of the course, when formative feedback was not largely available because it was impractical for a small team of instructors to consistently read and grade hundreds of short essays every week for 11 weeks throughout the semester.

The second kind of feedback was intended to be part of a larger goal in integrating meaningful feedback into the LMS. Because the essays were graded by machines throughout the semester, the instructors were able to shortlist a range of essays for deeper reading based on the scores that were generated with a level of confidence. With the ability to shortlist essays for further discussions with the students, they could gain a better grasp of the importance and significance of ethics and moral education based on their peers' work and through increased interactions with the instructors. By design, the formative feedback focused on argumentation, ethical reasoning, and critical analysis rather than looking at lower-level skills like grammar and sentence structure in the short essays. While the provision of this latter kind of feedback can be beneficial, especially when scaled with the university's learning systems, the focus in this study, however, was more on the former kind of feedback since it is critical to generate accurate results that allow the latter kind of feedback to emerge once the AES system stabilizes and performs robustly.

4. Findings and discussion

The findings mainly stem from phases two, three, and four of this study (from Figure 1). This section describes the accuracy of assignment coding, the training accuracy and validation of the model, followed by the implementation of the machine learning algorithm, and touches on the generated feedback as a result of the scoring.

4.1. Accuracy of assignment coding, training, and testing

During coding assignment of the training dataset, two raters initially coded a small subset of 20 short essays and provided reasons for the scoring of each essay, before coming together to resolve inter-rater differences. An inter-rater reliability (IRR) of 90.0% was initially obtained. The remaining differences were subsequently resolved after in-depth discussions between the two raters. After ensuring the two raters have achieved a high level of consistency with the scoring rubric, they proceeded to code another 167 short essays as part of the training set that was then used to train the model. The training accuracy from a training set of 187 short essays was 95.2%.

A 10-fold cross-validation was also conducted to evaluate the model and no overfitting was found. To test the model, 30 short essays scored by a human and the machine were compared. With the inclusion of the correction for chance agreement, the Kappa value for agreement between human raters and the machine model was 0.67, indicating substantial agreement between established labels by humans and the predictions by the algorithmic

model. These findings help to answer the research question that the model based on the Random Forest classifier in the AES system can perform similarly to a human rater, albeit with lower accuracy but with improvable performance, considering that this study consist of a single cycle of the scoring process and the algorithm's parameters can be further optimized.

4.2. Provision of formative feedback

The AES system presented scores for students' written essays during the study and although the scores were binary, they provided students with additional cues of how their work have been assessed and when used together with discussions on the online forums, students could better monitor and observe one's own activities, self-evaluate one's performance, and take actions based on the performance outcomes. These are important characteristics of self-regulated learning, important for academic performance as shown in other studies (e.g., Zimmerman & Moylan, 2009), but also form a critical part of how students engage themselves and others, with great relevance in the domain of ethics and moral reasoning.

The second form of feedback was given when the team of instructors provided comments on selected essays that they felt required a reaction or response. The reactions and responses may range from comments about well-written points or guidance on ideas and thoughts. By tapping on these examples, the general flow of ideas during the course can be better understood, similar to Lee and Tan's work (2017a; 2017b) and contribute towards a more productive discourse that benefits instructors and the students. If the essays were already deemed acceptable, no further comments were provided. Examples of the essays with respective scores and instructors' comments are shown in Table 3.

Table 3. Examples of essay excerpts with respective scores and formative feedback from instructors (if any)

Essay ID	Excerpt of essay on the topic of "utilitarianism" (word count)	Instructors' final label	Machine final score	Instructor comments
62	First and foremost, a utilitarian will have to consider the context in which why sex education is necessary before evaluating if he or she should proceed with it. Through sex education, students will learn important knowledge and insights into sex as a whole. The aim of the sex education in this case is not to reduce sexual behavior among students… but equip them with the knowledge to practice safe sex… (282 words)	1	1	Clear thesis with plausible reasons; good understanding of utilitarianism.
60	Firstly, a utilitarian would consider whether the decision of teaching sex education in a public school would be able to maximize overall well-being… For sensitivity context, certain aspects of the students such as their age and level of maturity should be taken into consideration before deciding whether sex education is suitable for them or rather what kind of sex education is more appropriate for that particular age group… (320 words)	0	0	Clear thesis but student appeals directly to claims about what is right and wrong, rather than deriving claims about right and wrong from the effects of actions on overall well-being.
121	From a utilitarian's point of view, their main aim is to maximize the well-being of the society. As such, we should take into account the possible benefits and implications of teaching sex education in public school. The main aim of sex education is to educate children on the potential issues related to sex. This is to prevent children from making wrong choices that may impact them greatly… (307 words)	1	1	[No comments from instructors]

4.3. Contributions and limitations of our study

In this study, our AES system has proved that it can automatically score and provide feedback to students of the "Ethics and Moral Reasoning" course with consideration of human factors. Although this capability has already

been demonstrated in some established systems (e.g., E-rater; Attali & Burstein, 2006) and emergent systems that use deep learning (e.g., Singla et al., 2021), these AES systems are also found to be over-stable (large essay changes cause little score variations) and over-sensitive (small essay changes cause large score variations). Our proposed HAI-influenced AES system partially negates these downsides with no overfitting,

However, a literal transfer of said system for use in other fields or algorithmic evolution into an all-encompassing type of algorithm with good accuracy is likely not possible soon. In other words, there is no one-size-fits-all algorithm that can be used without sacrificing certain aspects of accuracy, and although it is not an impossible task as demonstrated in an attempt by Olive et al. (2019), it can hardly compete with a predictive model for a specific course, such as the case in this study.

Nevertheless, it is possible to try and maintain a balance between achieving high accuracy (sensitivity and specificity) of essay scoring in a dedicated course and attempting to shift towards a slightly less accurate but general-use scoring system, particularly with human-based inputs and considerations. Although this effort will require tremendous resources to develop and maintain and likely not suit the objectives of every study or project, this limitation however will not detract from the benefits of developing AES systems for scoring large amounts of essays and HAI systems in general, allowing us to rethink and reflect on machine-based judgements.

5. Conclusions and future work

An automated essay scoring and feedback system was developed from open source to address several issues that arose from the running of an online course with large enrolments, further requiring automated assessment of students' work to better encourage meaningful teaching and learning. The study was divided into four phases and a mixed-method approach was used, with consideration of human-based inputs such as rubrics and qualitative coding of data subsets for training a machine learning model, which was then used to automatically score more essays at scale. Outputs and formative feedback in terms of essay scores and instructors' comments, which were lacking in previous iterations of the course, can then be provided to students, and possibly be used for fine-tuning the system's algorithms.

Returning to the research question, the AES and feedback system has shown to be beneficial in providing formative feedback to students, but it is still too early to decide whether this system can act as a surrogate for instructor interactions. This is because the implications and repercussions of replacing the teacher in a classroom can only be proven through multiple and sometimes longitudinal studies that provide evidence for explaining patterns of variables over time. However, it is undeniable that having an existing automated system that analyses and scores student essays does ease the load of instructors and provides instructors with more time to enhance activities in the course while gaining the ability to measure learning gains when needed, a benefit from the implementation of HAI design that considers human conditions and contexts.

As part of future work, once the AES and feedback system has been made more reliable and robust after several runs and validation processes, it can be integrated with an existing LMS to answer other interesting research questions, such as: "How much human interaction is required for students to feel their instructors are academically invested in them?" and "do students that receive automated feedback improve the quality of argumentation and decisions more than students who do not receive feedback?" These research questions will help drive vested interests to achieve "specific, measurable, agreed upon, realistic, and time-based" goals of smart AI research (Yang, 2019), that are generic enough to be understood by the public and also with wide-ranging implications that are meaningful to the masses.

Automated essay scoring, as a vital machine learning application over the last few decades, remains important to both instructors and students in providing summative and formative feedback for improving teaching and learning. The recent introduction of human-centric factors and adjustments to AES systems, however, has greatly helped to make learning visible and relevant to emergent user needs. As the need for AES becomes more imperative with the growing emphasis on remote and online learning, and with the aid of emerging techniques and technological affordances, the use of HAI designs in automated essay scoring may eventually become more widely implemented and commonplace.

Acknowledgement

This study was funded by the Startup Fund under the Centre for Research and Development in Learning (CRADLE), Nanyang Technological University, Singapore. The views expressed in this paper are the authors' and do not necessarily represent the views of the host institution. The research team would also like to thank the instructors and student participants involved in this study.

Conflict of interests

The authors confirm there are no conflicts of interests.

References

Aoun, J. E. (2017). *Robot-proof: Higher education in the age of artificial intelligence.* MIT press.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment, 4*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cohen, J. (1960). A Coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. https://doi.org/10.1177%2F001316446002000104

Cope, B., Kalantzis, M., & Searsmith, D. (2020). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory*, 1-17. https://doi.org/10.1080/00131857.2020.1728732

D'Mello, S. (2021, July 1). *From modeling individuals to groups: It's a multimodal multiparty* [Keynote session]. Educational Data Mining 2021, Paris, France.

Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A Cross-disciplinary Perspective* (pp. 71-86). Routledge.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2). http://imej.wfu.edu/articles/1999/2/04/index.asp

Garcia-Magarino, I., Muttukrishnan, R., & Lloret, J. (2019). Human-centric AI for trustworthy IoT systems with explainable multilayer perceptrons. *IEEE Access, 7*, 125562-125574.

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: "Breakthrough? Or buncombe and ballyhoo?". *Journal of Computer Assisted Learning, 37*(5), 1207-1216. https://doi.org/10.1111/jcal.12577

Ghanta, H. (2019). *Automated essay evaluation using natural language processing and machine learning* (Unpublished master's thesis). Columbus State University, Columbus, GA. https://csuepress.columbusstate.edu/theses_dissertations/327/

Hattie, J. (2013). *Visible learning: A Synthesis of over 800 meta-analyses relating to achievement.* Routledge.

Hattie, J., & Timperley, H. (2007). The Power of feedback. *Review of educational research, 77*(1), 81-112. https://doi.org/10.3102%2F003465430298487

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education*. Center for Curriculum Redesign.

Juwah, C., Macfarlane-Dick, D., Matthew, B., Nicol, D., Ross, D., & Smith, B. (2004). Enhancing student learning through effective formative feedback. *The Higher Education Academy, 140*, 1-40.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An Introduction to latent semantic analysis. *Discourse processes, 25*(2-3), 259-284. https://doi.org/10.1080/01638539809545028

Lee, A. V. Y. (2020). Artificial intelligence in education (AIEd). In H.-J. So, M. M. Rodrigo, J. Mason, & A. Mitrovic (Eds.), *Proceedings of the 28th International Conference on Computers in Education (ICCE), Volume 2* (pp. 749-750). Asia-Pacific Society for Computers in Education.

Lee, A. V. Y. (2021). Determining quality and distribution of ideas in online classroom talk using learning analytics and machine learning. *Educational Technology & Society, 24*(1), 236-249.

Lee, A. V. Y., & Tan, S. C. (2017a). Discovering dynamics of an idea pipeline: Understanding idea development within a knowledge building discourse. In W. Chen, J.-C. Yang, A. F. Mohd Ayub, S. L. Wong, & A. Mitrovic (Eds.), *Proceedings of the 25th International Conference on Computers in Education (ICCE) 2017* (pp. 119-128). Asia-Pacific Society for Computers in Education.

Lee, A. V. Y., & Tan, S. C. (2017b). Understanding idea flow: Applying learning analytics in discourse. *Learning: Research and Practice, 3*(1), 12-29. http://dx.doi.org/10.1080/23735082.2017.1283437

Lepri, B., Oliver, N., & Pentland, A. (2021). Ethical machines: The Human-centric use of artificial intelligence. *IScience, 24*(3), 102249. https://doi.org/10.1016/j.isci.2021.102249

Liu, N. F., & Carless, D. (2006). Peer feedback: The Learning element of peer assessment. *Teaching in Higher education, 11*(3), 279-290. https://doi.org/10.1080/13562510600680582

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In *International Conference on Information Computing and Applications* (pp. 246-252). Springer.

McCarthy, J. (1995). What has AI in Common with philosophy? In *International Joint Conference on Artificial Intelligence* (IJCAI) (pp. 2041-2044).

McKeachie, W. J., & Svinicki, M. (2006). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers.* Houghton Mifflin Company.

Nguyen, T. T., & Walker, M. (2016). Sustainable assessment for lifelong learning. *Assessment & Evaluation in Higher Education, 41*(1), 97-111. https://doi.org/10.1080/02602938.2014.985632

Norvig, P. (2007). *How to write a spelling corrector*. http://norvig.com/spell-correct.html

Olive, D. M., Huynh, D. Q., Reynolds, M., Dougiamas, M., & Wiese, D. (2019). A Quest for a one-size-fits-all neural network: Early prediction of students at risk in online courses. *IEEE Transactions on Learning Technologies, 12*(2), 171-183. https://doi.org/10.1109/TLT.2019.2911068

Page, E. B. (1966). The Imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238–243.

Rest, J. R., Thoma, S. J., & Bebeau, M. J. (1999). *Postconventional moral thinking: A Neo-Kohlbergian approach.* Psychology Press.

Rokade, A., Patil, B., Rajani, S., Revandkar, S., & Shedge, R. (2018). Automated grading system using natural language processing. In *2018 Second International Conference on Inventive Communication and Computational Technologies* (ICICCT) (pp. 1123-1127). IEEE. https://doi.org/10.1109/ICICCT.2018.8473170

Roselli, R. J., & Brophy, S. P. (2006). Experiences with formative assessment in engineering classrooms. *Journal of Engineering Education, 95*(4), 325-333. https://doi.org/10.1002/j.2168-9830.2006.tb00907.x

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment, 1*(2). https://ejournals.bc.edu/index.php/jtla/article/view/1668

Sareen, S., Saltelli, A., & Rommetveit, K. (2020). Ethics of quantification: Illumination, obfuscation and performative legitimation. *Palgrave Communications, 6*, 20. https://doi.org/10.1057/s41599-020-0396-5

Schonlau, M., & Zou, R. Y. (2020). The Random forest algorithm for statistical learning. *The Stata Journal, 20*(1), 3-29. https://doi.org/10.1177%2F1536867X20909688

Shavelson, R. J., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., Tomita, M. K., & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A Collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*(4), 295-314. https://doi.org/10.1080/08957340802347647

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. https://doi.org/10.3102%2F0034654307313795

Singla, Y. K., Parekh, S., Singh, S., Li, J. J., Shah, R. R., & Chen, C. (2021). *AES systems are both overstable and oversensitive: Explaining why and proposing defenses*. PsyArXiv. https://doi.org/10.48550/arXiv.2109.11728

Taghipour, K., & Ng, H. T. (2016). A Neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882-1891). Association for Computational Linguistics.

Thille, C., Schneider, E., Kizilcec, R. F., Piech, C., Halawa, S. A., & Greene, D. K. (2014). The Future of data-enriched assessment. *Research & Practice in Assessment, 9*, 5-16.

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An Overview of current research on automated essay grading. *Journal of Information Technology Education: Research, 2*(1), 319-330. https://www.learntechlib.org/p/111481/

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmar, M., & Nerini, F. F. (2020). The Role of artificial intelligence in achieving the sustainable development goals. *Nature Communication, 11*(233), 1-10. https://doi.org/10.1038/s41467-019-14108-y

Vozzola, E. (2014). *Moral development: Theory and applications.* Routledge.

Walker, E., Wong, A., Fialko, S., Restrepo, M. A., & Glenberg, A. M. (2017). EMBRACE: Applying cognitive tutor principles to reading comprehension. In *International Conference on Artificial Intelligence in Education* (pp. 578-581). Springer, Cham.

Yang, S. J. H. (2019, December 2-6). Precision education: New challenges for AI in education [Conference keynote]. In *Proceedings of the 27th International Conference on Computers in Education (ICCE)* (pp. XXVII-XXVIII). Asia-Pacific Society for Computers in Education.

Yang, S. J. H. (2021). Guest Editorial: Precision education – A New challenge for AI in education. *Educational Technology & Society, 24*(1), 105-108.

Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence, 2*, 100008. https://doi.org/10.1016/j.caeai.2021.100008

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In *Handbook of Metacognition in Education* (pp. 311-328). Routledge.