

Values & Ethics in AI
Final Project: AI-Enabled Grading Systems
John Ware
10 December 2022

Introduction

Artificial intelligence is making its way into the classroom with AI-enabled grading systems. From smart tutoring systems to autonomous grading, AI has the potential to revolutionize our education system and greatly improve the student experience. New Educational Technology (EdTech) applications advertise attractive features like instantaneous feedback, customized lesson plans, and personalized tutoring services built with state of the art natural language processing, computer vision, and machine learning tools.

Perhaps one of the most hotly debated applications has been the use of AI for grading and student evaluation. AI and software systems have been used for a wide range of evaluation techniques, ranging from the rather innocuous example of e-learning recommender systems for student exercises, up to qualitative assessments of long-form essay responses. These systems can significantly reduce evaluation time for teachers and put feedback in the hands of students near-instantaneously, but these features come at a high cost. These systems, as with many current AI systems, are subject to algorithmic bias, misuse, and significant inaccuracies, all of which have the potential to negatively impact student progress and trust in academic institutions.

AI-enabled grading software helps educators autonomously grade student work and provide varying levels of feedback against metrics set by individual teachers. These types of systems have been in use at varying levels for multiple decades, but their applications have gone largely unnoticed or unrecognized by the general public. Advancements in machine learning, computer vision, and natural language processing have all contributed to the rise of these software systems, making the technology accessible to a variety of educational institutions across the U.S. and world. Recent improvements in computer vision and natural language processing in particular have caused an explosion of use cases and applications across the education sector, raising new concerns and debates about the ethics and accuracy of these grading systems.

This paper will explore the topic of AI-enabled automated grading systems through the lens of explainability and transparency, and address the following questions:

- What types of automated grading systems exist, and how are they currently used?
- How can educators ensure these grading systems and their results are explainable, and transparent to the greatest extent possible?
- What ethics and values should be considered in the deployment of these systems?

- How can these systems be used to provide adequate and appropriate feedback to students?

Grading Implementations & How They Work

While many new grading systems claim to be AI-based or AI-enabled, the main differences between them can be found in the subdisciplines of AI they use to assess performance. Figure 1 shows a breakdown of several subsets of Artificial Intelligence, many of which are being applied to educational applications today.

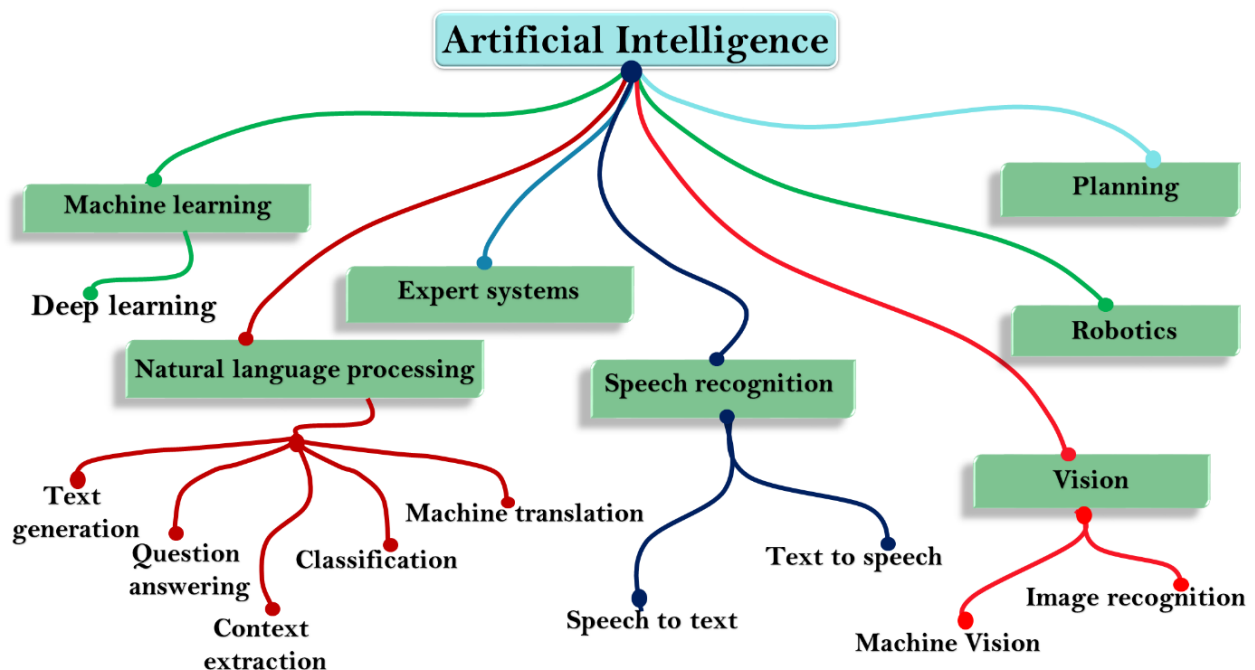


Figure 1. Artificial Intelligence Disciplines¹

Rule-Based

Perhaps the simplest implementation is a rule-based artificial intelligence, whereby a system grades performance based on a set of predetermined criteria or rules. This could look like a list of targeted keywords in a short answer response, or a set of mathematical terms in a formula; if a student response contains these predefined targets, they achieve a certain score. This type of automated grading is the most common in educational settings, and is commonly found in applications like Blackboard and Canvas. These systems benefit from being very explainable and transparent; the outcomes from these systems can be understood by students and teachers, and they can be challenged and updated as necessary when a mistake is made. These systems therefore work well for consequential

¹ <https://www.javatpoint.com/subsets-of-ai>

evaluations, though their applications are limited to assessments with predetermined answers such as textbook and information recall problems. Their primary benefits over human educators is the speed at which they can provide an assessment, and the scale at which they can be deployed.

Machine Learning

A step up in complexity from these rule-based structures is a machine learning based grading system. Machine learning models can be built from past performance data, including textual information. They can be deployed to evaluate text or numerical inputs based on a combination of rules and past data trends, and they can be used to predict performance or identify when new data is out of the ordinary. In the case of essay analysis, machine learning models can evaluate a large dataset of past writing samples with human-provided feedback, and “learn” from the trends and techniques seen in those samples. The model can then apply similar feedback on new writing samples, and do so at a scale and speed far exceeding that of human educators. These models are only as good as the training data they use, meaning they cannot interpret information that has not been seen in prior writing samples, and they cannot provide feedback or criticism that has not already been provided in another setting. Machine learning systems are notoriously non-transparent, and they cannot provide direct traceability to the reasons they provided a specific instance of feedback. This sort of implementation can work well for essays that are based on common, reused prompts with a large historic dataset from which the model can train; in these cases, the algorithm could provide the same grade distribution as a human, but not be able to explain it to a student. Because they cannot provide transparency nor explainability, these systems have long struggled to gain the trust of students and educators, despite replicating the statistical results of human graders in most scenarios². While these types of systems cannot compete with high-quality human one-on-one grading and feedback, they can greatly increase efficiency at scale for a subset of essay grading scenarios, and thus have found widespread applications in standardized testing³.

Machine learning algorithms find patterns that are imperceptible to humans, but they are nonetheless statistical patterns; algorithms can identify data features that are highly correlated with a dependent variable such as an assignment grade. These models cannot evaluate intangible features such as creativity or clarity, as they are only processing quantitative representations of the data. When these systems are provided new writing samples to grade, they prioritize the features that are most correlated with an overall grade. This has been shown to inadvertently incentivize certain features over others, such as keywords and essay length, which can be statistical indicators of a high grade. If we assume there is no change to student inputs in the future, these models would continue to follow the statistical trends, and likely continue to outperform human graders at scale.

² <https://www.thenewatlantis.com/publications/machine-grading-and-moral-learning>

³ <https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays>

However, if the key features were to be identified and exploited for a given algorithm, students could essentially “game” the system to achieve better grades with anomalous inputs like long incoherent responses with a set of keywords.

This concern is not necessarily unique to machine learning though; it was also raised for SAT essay response scores before machine learning was used, and it was found that word count and sentence structure outweighed accuracy and correctness⁴ even for human graders. While the application of machine learning could help alleviate these concerns in some ways, a statistical model cannot fully eliminate the potential for exploitation. It also cannot fully address the desire for explainability, as these systems cannot provide detailed or comprehensive feedback for writing samples. At best, they can only provide quantitative metrics related to data features they can measure (i.e. word count and number of keywords found), so a student claiming they answered the prompt in an “accurate and succinct way” has little or no explanation or recourse for their grade.

Computer Vision

Another layer of artificial intelligence involves computer vision, whereby a computer is trained to “view” an image in an attempt to mimic human perception. This has recently grown in popularity, in part because of its ability to address explainability concerns. Computer vision algorithms can process an image and identify individual subfeatures, then compare those subfeatures to a predefined rubric. Figure 2 shows one such example of this technology with a software system known as Gradescope.

Gradescope is a grading software that streamlines the grading process for teachers⁵; the company claims that grading time can be reduced by up to 50%. The software is semi-autonomous, and allows for teachers to create custom rubrics for individual questions. It takes the inputs from each student, and groups responses by question so the teacher can assign full or partial credit to a group of similar answers. This works exceptionally well with visual or short form verbal responses where specific answers are necessary, such as grade school quizzes or science/engineering-based courses. Given the similarity in responses, the software can use tools like computer vision to identify specific segments of a student response, and compare those against a custom rubric that has been created by the teacher.

⁴ <https://www.nytimes.com/2005/05/04/education/sat-essay-test-rewards-length-and-ignores-errors.html>

⁵ <https://www.uml.edu/it/services/academic-technology/grading-software-gradescope.aspx>

The screenshot displays the Gradescope interface for a course titled 'Object Oriented Software Engineering'. The assignment is '2: Git and Requirements', which is marked as 'GRADED'. The student's overall score is 25.5 / 35 pts. The specific question being evaluated is 'git commit graphs', with a score of 5.5 / 15 pts. The student's submission is a hand-drawn Git commit graph. The evaluation on the right side of the interface lists various features and their scores:

- 0 pts Correct
- 0.5 pts Did not move HEAD after checkout
- 0 pts Tag should be on a commit, not another tag
- 10 pts Missing Part 1A
- 0 pts **Regular Merge and not a Fast Forward Merge**
- 1 pts Branches (such as "master", "cookies", ...) should be represented as pointers to a specific object
- 9 pts Used git log graph instead of drawing
- 5 pts Missing part 1B
- 0.5 pts Missing a commit
- 0.5 pts **Missing a branch pointer**
- 0.5 pts Missing a step
- 1 pts Branch tags should be removed after branch deletion
- 1 pts Missing: "bread" merge produces a merge commit

Figure 2. Gradescope Example Evaluation⁶

In the example shown in Figure 2, a rubric has been set by a professor to include a list of features necessary for a perfect score (shown on the right of the image). The algorithms are able to detect individual features from a student's diagram and labeling (such as flow charts or formulas), and compare those features against the rubric to assign a score. This score is clearly recorded and broken down for the student, which makes the software both explainable and transparent (to a certain extent). This example of software also has a regrade request feature whereby a student can request a human review of an image or label if they feel something has been improperly assessed; this serves to both build confidence from students and teachers, and improve the AI algorithms by providing additional labeled training data.

⁶ <https://ii.library.jhu.edu/2018/02/15/grading-in-the-fast-lane-with-gradescope/>

Computer vision products like this can also be used to interpret written text, and translate that into a digital form from an image. While handwriting can be difficult to interpret and these systems are not perfect, they are able to take a block of text and compare it against a set of keywords or phrases that are also set by an educator. In this way, one can combine the principles behind rule-based AI and computer vision AI to evaluate a short form response based on a set of known answers. This provides the benefits of both approaches - evaluation at scale from a variety of inputs, and a known set of rules for a transparent and explainable grade that is not achievable with basic machine learning or statistical models.

Natural Language Processing

In recent years, natural language processing (NLP) has seen several significant advancements and further adoption in the tech industry. This builds on the basic text analysis tools that are paired with machine learning, and recognizes associations between words and phrases used in regular dialogue and prose. These NLP engines are trained on a massive amount of content, and are often built with convolutional neural networks, making them neither explainable nor transparent. These toolkits are currently the closest approximation to natural language in an artificial form, with examples including OpenAI's GPT-3⁷ and Python's Natural Language Toolkit (NLTK)⁸.

For prompts that require creative and unique responses that cannot be predicted ahead of time, a natural language tool like these could feasibly be used to assess a writing sample against a list of intangible goals. In a perfect future scenario, a teacher or educator could input a list of values such as "depth of reasoning" or "applies a variety of course content", and a natural language tool such as GPT-3 could both accurately interpret those goals and evaluate a writing sample against them. In this scenario the evaluation would match or exceed the quality of a human evaluation (given the engine's depth of training), and be capable of deploying at a much larger scale to provide near-instantaneous feedback to students.

In reality, none of the current natural language toolsets are perfect, nor can they perform at the levels described above. GPT-3 is often categorized as the largest and most advanced natural language toolkit currently in use⁹, and yet it is not capable of complex long-form responses or evaluations on its own. OpenAI has provided an open source playground for GPT-3¹⁰, and anyone with an account can quickly explore the strengths and limitations of the engine. After a few iterations, it is relatively easy to see that your interactions with the computer are not the same quality as a human interaction, and responses from GPT-3 somewhat mimic those expected of a child. While this is an intriguing capability, and

⁷ <https://openai.com/api/>

⁸ <https://www.nltk.org/>

⁹ <https://towardsdatascience.com/gpt-3-a-complete-overview-190232eb25fd>

¹⁰ <https://beta.openai.com/playground>

certainly a step in the right technological direction, it does raise concerns about applying this sort of tech to consequential grading in an educational setting.

Vice News conducted a survey of automated essay scoring using AI engines across the United States, and found that at least 21 states use natural language processing-based AI systems for automated grading¹¹ as of 2019. Figure 3 shows a map of the U.S. with states using AI-grading systems colored in dark blue.

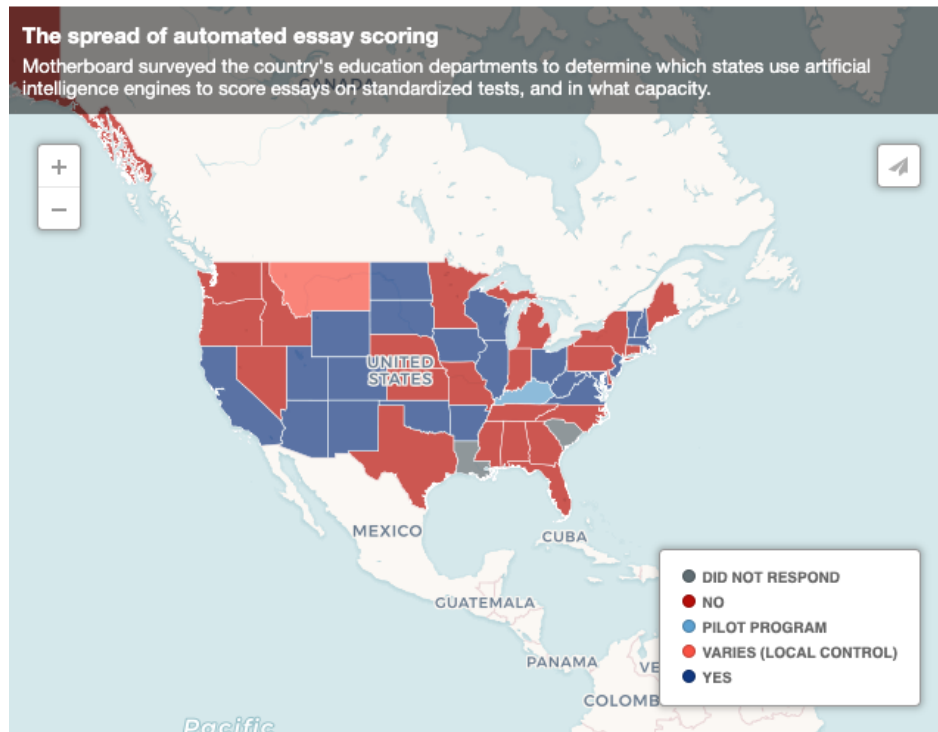


Figure 3. Automated Essay Scoring in the U.S.¹²

Their investigation into an Educational Testing Service algorithm showed that writing scores generated from the NLP-enabled engine were closely correlated with features like essay length and grammar (i.e. keyword usage). When compared with human grading, it was found the NLP-enabled algorithm scores differed greatly along racial and cultural lines, meaning scores from students with different language backgrounds were either improved or worsened from the automated scoring system depending on where they were from and what language they originally spoke¹³.

Similar to the simpler machine learning models applied to textual data, these systems cannot provide detailed explanations for the scores they provide, nor can they provide

¹¹ <https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays>

¹² https://tosfeathers.carto.com/viz/59accc81-970c-4768-bc64-7582658b001b/public_map

¹³ <https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays>

transparency into their “decision” process. At best, they can provide quantitative measures for a set of defined features associated with the writing sample (i.e. word count, number of keywords, etc.), but the existence of these features is purely based on the training dataset used to create the natural language tools. In other words, these NLP-engines do not seem to truly understand the information they are processing as a human would, but rather they excel at demonstrating passable word association for complex topics.

Importance of Explainability

To understand the importance of explainability for these grading systems, we must first understand the importance of providing feedback to students. Students learn by trial and error; they continuously perform increasingly difficult tasks, improving primarily through making mistakes and adjusting course in response to feedback (in a variety of forms). A good educator generally strives to provide meaningful and constructive feedback to students on their work, which is critical for their progression and growth.

Several studies and general teaching practice has shown that effective feedback for students must be timely¹⁴; it needs to be presented to the student as quickly as possible after the exercise or assignment is complete. Feedback also needs to be educational and clear; a student must be able to understand what they did well, what they need to improve upon, and why they earned a certain grade. In other words, ideal feedback needs to be near-instantaneous and explainable.

Explainability should therefore be a priority and critical metric prior to the deployment of any of these technologies into an educational setting. For any assignments where an AI-enabled grading system replaces a human grader, the level of explainability should always be equal or better to that of a human.

In theory, and somewhat in practice, these AI-enabled grading systems can provide this ideal feedback. In many cases, AI-enabled evaluation systems can far outperform human grading in the areas of timeliness and explainability, especially for assignments where a particular answer is sought. In these scenarios, the automated systems can provide instant feedback, clearly marking both right and wrong elements of a student response, as shown in the Gradescope example.

These discussions apply to teachers and students alike, along with anyone working in educational institutions. Every student and learner needs to have a deep understanding of how they are scored on each assignment and how they could improve. Instantaneous or

14

https://sc.edu/about/offices_and_divisions/cte/teaching_resources/grading_assessment_toolbox/providing_meaningful_student_feedback/index.php

timely feedback is critical for improvement, especially in the case where topics build on one another. With high student to teacher ratios and complex assignments, it can be difficult for teachers to keep up with the demand for this feedback before the next topic overwhelms a student, creating a cascading negative effect on performance. As AI-enabled systems are introduced, we must take care to not replace one risk with another, potentially greater one.

Educators need to be able to set goals and performance measures for these grading systems, and have an understanding of how grades are allocated and distributed. They must also understand how these assigned grades compare to a control set of human-graded material, and continuously re-evaluate this as new content is added. This can help teachers address common fault lines, and adjust course as needed throughout a class. Educators must be able to translate grades into feedback, and vice versa, in order to provide the necessary actionable information to students. When automated grading systems are applied, educators must be able to see explainable results and ensure proper traceability to correct responses and grades.

Evaluators of these students and educators (i.e. college admissions, school administrators) must be able to understand the extent to which individuals are relying upon this automated grading software, and ensure the quality of education is upheld above all else. They must be able to understand the limits of these systems, and see explainable and transparent results so they can incorporate additional information in their review (such as human-graded material, extracurricular activities, etc.).

Challenges with Deployment

Explainability can be achieved in a variety of ways with each of the algorithmic approaches defined above. In a technical sense, language processing kits could provide a sample of writing comparisons for each grade level to help students develop an intuitive sense for writing quality; this is especially important when specific quantitative feedback is not feasible.

In a practical sense, natural language processing kits could use conversational language to help a student understand where they should focus on improvements. Feedback phrases like “improve your grammar” or “expand your vocabulary” could be provided as actionable information for the student. In the future, these NLP engines could begin to learn about the individual learning styles for each student, and tailor their feedback to each student. There is an essential human element for qualitative responses, where an educator must weigh a variety of factors to score students and provide feedback in a way that builds the student without discouraging them. For qualitative or creative assessments to match those of a high quality human grader, the system must be able to predict and interpret a student response to given feedback.

Current NLP engines like GPT-3 are able to generate content in response to certain prompts far faster and easier than they can evaluate an individual's writing. There are also commercial products and resources built on the GPT-3 library to help enable unique content generation¹⁵. Some services like Jasper boast subscription services for their AI engine to write full-length marketing and blogging content with minimal user input, all from a user-friendly API as shown in Figure 4. While this particular service is geared towards business marketing tasks, one could envision applying this capability to academic prompts whereby students can submit the work as their own.

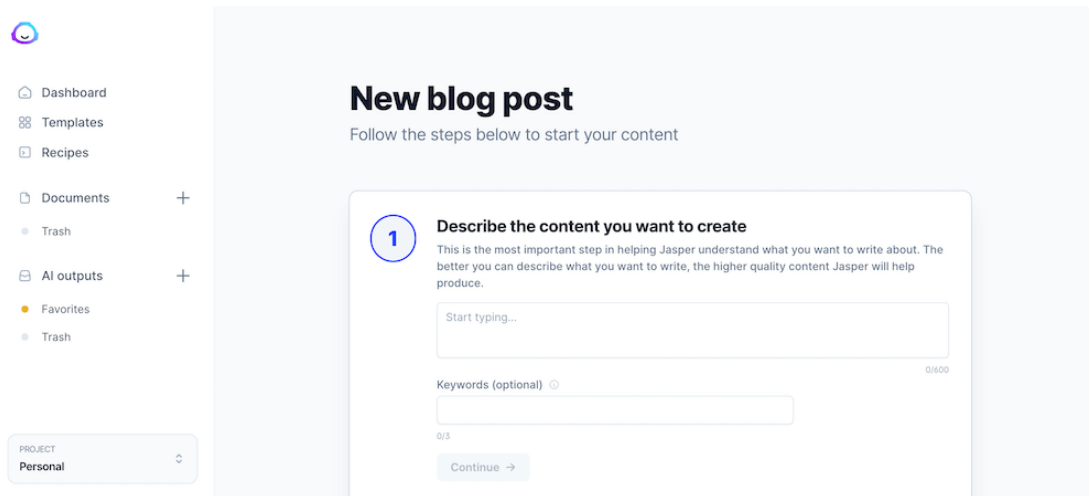


Figure 4. Jasper API¹⁶

This sort of capability creates a new set of challenges when applied to education, as students can use these tools to generate unique responses to essay prompts without writing themselves, and educators will have a difficult time identifying or addressing this “fake” content. This can result in widespread cheating before educators have the resources or knowledge to stop it.

Another potential challenge in the deployment of these systems is the management or mitigation of false positives and false negatives: instances where a machine assessment differs greatly from a human assessment. While these rates may be low overall, each instance has the potential to have significant psychological effects on an individual, and may greatly affect their ability to progress in their education. With human feedback, students are able to reconcile their emotions and discuss ways to improve with teachers, but this may not hold true with initial automated grading systems.

¹⁵ <https://openai.com/blog/customized-gpt-3/>

¹⁶ <https://www.jasper.ai/>

These same false positives and negatives can degrade trust in the algorithms, from teachers and students alike. This can lead to algorithm aversion, which can be a difficult force to overcome in the deployment of new technology. These systems will almost certainly have flaws, and it will be important to continually assess their performance against human grading. Distrust in algorithms can be further fueled by systems that offer no reconciliation process for students¹⁷, which contributes to a lack of explainability and greatly disheartens the affected students.

Perhaps the greatest challenge in the deployment of these systems manifests in unique open-ended prompts with creative and unusual responses. Out-of-the-box thinking is not well understood by the current state-of-the-art natural language processing tools, and these grading systems struggle to address responses that stray from standard formulaic structure. Students would have minimal recourse to address their grade in these instances, so a robust human-regrade process must be included.

The area of technology policy also presents challenges for these grading systems. As with other technological advancements, national policy is not matching the pace of change seen in automated grading. Without national or local policy around the technology and its applications, local school boards and institutions will shoulder the burden of evaluating, procuring, and deploying these capabilities. While this does offer a great deal of freedom and flexibility for these institutions and their particular needs, it raises ethical questions around how individual AI-based products are chosen and deployed. Inconsistent rollout can create and fuel educational disparities across the country, and the direction of those disparities is yet to be determined.

Conclusion & Recommendations

This technology has the potential to have groundbreaking effects on education and the way we evaluate student performance. In an ideal state, these AI-enabled automated grading systems could provide students with instantaneous and meaningful feedback, with the same or better quality as a human grader. These systems could be applied (in the future) to any type of assignment, be it a quantitative math problem or a qualitative and creative essay prompt. They could even serve as tools with which we can adjust our style of assessment altogether; so-called “stealth assessments” using these types of tools have shown promising results in reducing test anxiety and providing a means of ongoing evaluation¹⁸.

Despite this positive potential, the current state-of-the-art autonomous grading systems are nowhere near this level of quality or preparedness. While each of these AI-related technologies can help improve student grading, they also come with a list of upsides and

¹⁷ <https://hbr.org/2020/08/what-happens-when-ai-is-used-to-set-grades>

¹⁸ https://myweb.fsu.edu/vshute/pdf/Stealth_Assessment.pdf

downsides, some with a greater significance than others. At a minimum, educators risk providing students grades without feedback, which can have devastating effects on a student's progression. At worst, a poor deployment of these systems can have outsized effects on a student's psychological and physical progression through school.

In evaluating these AI-enabled grading systems, educators must prioritize explainability and transparency. They must also consider and balance the values of fairness, self-determination, and net benefit. Any chosen system should provide fair assessments to every student, regardless of background or past performance. Educators must also not lose sight of individual students, even in the event these systems provide an overall net benefit across a student population; educators must always strive to help every student improve. It is critical that students are encouraged and enabled to take agency in their own education, and educators must ensure these automated systems do not outweigh or overrule human evaluation.

Based on the current state of this technology, it is recommended that these automated grading systems should not be used in high-stakes or consequential scenarios where student advancement is on the line, such as in major national testing or college admissions. Any deployment of these systems should have strict requirements for transparency and feedback, and blind scoring and grading should be rejected. Fair use principles and guidelines should be coordinated amongst school boards and national educational institutions to specify the appropriate use-cases and rollout plans for each new autonomous grading system. Overall, we must continue to push forward with the development and improvement of these systems, and take great care to ensure early failures do not prevent long term success.