Criteria, Judgements and Automated Marking: Return to Lyotard

Introduction:

Automation is expanding into more and more areas of human activity. If done well, automation frees up individuals from mundane and repetitive labour, leaving more time for more advanced tasks or recreational activities. There is little doubt that automation of manual labour is beneficial, but what happens when intellectual labour becomes substituted by automation as well? Over 800 articles from Washington Post were written by bots in 2016, Wall Street Journal writes its routine stories about the state of the stock market, NLP applications in healthcare, not to mention the myriad of decentralised and covert uses of automated bots and scams. It is no surprise, then, that automation is trying to find its way into education as well. Some of the automation is relatively uncontroversial, such as tracking the attendance or scheduling. Automated marking is quite harmless in cases of tests with multiple choice answers, because the criteria for correct and incorrect answers is clearly established. However, over the last few years automated marking has begun to move into marking of essays at highschool and college level education. At this level, the criteria for judgement becomes a central question when designing the automated applications and making sure they judge well. In this paper I will consider the possibility of automated algorithms marking a 2000 word essay,which is what is written by most undergraduate students in Newcastle University. After a brief consideration of the technical side, the paper will focus on what it means to judge in the context of marking.

Consider marking a very good essay written by a student in your course. There are a set of rubrics that are meant to delineate the criteria for evaluating their work such as knowledge and academic skills, rationalisation and argument, and execution. They are then further separated into smaller, more precise rubrics such as subject competence or understanding of question. However, the rubrics for those judgements are always vague and open to interpretation. For ‘coherence and structure’ a great student is expected to “Combine different ideas together effectively and establish original linkages. To get above 70 in subject competence, the student must “Demonstrate excellent discipline knowledge and ability.”, but there is no further criteria for how to judge excellence. It remains vague. In fact, as this paper will argue, it must remain vague, with only broad rubrics for marking presented and without a specific set of criteria. The reasons for why this should be the case will be explained by relying on Lyotard and ultimately circle back to why no automated system available at the moment (or indeed in the foreseeable future) can adequately mark students' essays.

Return to Lyotard:

Here, guided by the ideal of education as a creative endeavour, we can agree with Lyotard that there is not, or at least, should not, be any universal criteria by which we judge the success of education. To clarify, this paper makes a distinction between criteria and reasons for judgement. Criteria is a universal rule, by which any judgement within a set framework can be made. Reasons for judgement are non-universal arguments for a particular judgement. They may exist throughout a given framework, but are not universal in the way criteria are.

As I will argue, Lyotard's predictions about the state of education have come true - education is now a productive endeavour, focused on performativity and efficiency. There is no emancipation to be had in education anymore and by majority it is seen as simply the pragmatic stepping stone to finding a job. While lamenting of this attitude may be misplaced, as such an approach to education has arguably been the case since industrialisation, it provides us with an ideal of what education could be, what it strives to be if only theoretically.

According to Lyotard, we should be like pagans and go beyond simply adapting our judgements to a given situation, but actually change the criteria by which we judge accordingly. Furthermore, these criteria can never be fully articulated. This is a necessary premise to the argument put forward in this essay - automation of marking (in other words, judging) in higher education will always fall short, because the criteria by which students are judged cannot ever be fully expressed. To explain this view further I will rely on an idea central to Lyotard's philosophy - the language games. First the paper will look at a couple of different ways that the marking is being automated.

About automation of marking in education:

The companies that provide the service of automated marking keep the source code of their programs private, so it is up to the researchers to figure out the mechanism employed. In the models with limited use of deep neural networks, the set of criteria by which the algorithms judge a piece of writing can be determined and split into a set of categories. One of the most popular applications, E-Rater uses Grammar, Usage, Mechanics, Style and Organisation & Development as its criteria. When looked under the hood, however, these criteria appear to be misguided. Perhaps it will suffice to say that several researchers have shown that the category of Organisation & Development is closely correlated with the essay, paragraph, sentence and word length. The more long words a student uses, the better marks they are likely to get. Even if they don't understand what the words mean.

As shown in a recent paper called "Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'" by John Gardner et. al.[1], the elements that researchers thought had to be considered such as grammar, style and mechanics have now evolved to require much more sophisticated elements to be evaluated such as curiosity, openness, engagement, creativity and persistence. What this shows is not simply the irreducibility of complex essay writing. Some degree of reducibility is quite easily achieved and has been done for a few decades now. The difficulty is in finding the criteria according to which the evaluation should be reduced to. As will be shown via the use of Lyotard, such criteria is impossible to deduce. Arguably, every attempt at such reduction either ends in failure or develops into a new language game.

[1] Gardner, J., O'Leary, M. and Yuan, L. (2021).

When the attempts to make essay marking automated began, the set of rubrics by which these judgements were supposed to be made were relatively straightforward and simple. As the computational power increases, the potential for different sets of calculations increases as well. And once Big Data became viable as datasets for deep learning algorithms, the criteria as such quickly vanished into what is referred to as the “black box”. In this context the “black box” refers to the impossibility of knowing what criteria the application uses to evaluate its function. In other words, the algorithm marks the essay, but no person knows the criteria by which the judgement was made exactly.[2]

Problems:

Les Perelman is perhaps the most well known critic of automated marking systems. He, together with some undergraduate students created BABEL, a program designed to write essays that produce gibberish but score high on the automated marking systems. Perelman has also identified that the criteria used to generate marks correlates highly with the length of the essay, providing further evidence that algorithms, despite giving fancy names for their reduction, have essentially been reducing the criteria to the lowest common denominator - length.

In 2018 researchers created a tool that could detect BABEL’s gibberish and so E-rater could no longer be fooled by the application. As Perelman notes in his new paper: ”This effort, however, solves a problem that does not exist”[3]. Ultimately, the students will learn to exploit elements of BABEL’s tricks. Using automated marking makes students better test takers, not better writers. To better illuminate the problem, the paper now moves onto Lyotard and the idea of judgement without criteria.

Towards Judgement Without Criteria:

First, a brief explanation of what judgement without criteria means. As mentioned before, criteria for Lyotard means a certain presupposed universality behind a set of judgements. A rule that applies in all language games, in all situations. Something that is rigid and unchanging. To judge without criteria is to judge without an assumed universality of that judgement. It is not, however, to judge without reasons. One way to interpret this is to understand judgement without criteria as a judgement whose criteria cannot be articulated.

Paganism:

In *The Postmodern Condition* Lyotard rejects metanarratives and universal truths and claims that everything is driven by narratives. However, he is not a pure relativist believing that anything goes. Which leads to a problem: Lyotard must determine how to form judgements without relying on universal criteria.

Our point of departure is to firstly consider Lyotard’s paganism. Paganism is of course meant as a metaphor. Lyotard is looking for ways to consider justice in a (godless) pagan society,

[2] Examples of such systems can be found here: https://aclanthology.org/D16-1193, https://link.springer.com/chapter/10.1007/978-3-319-99722-3_18 .

[3] Perelman, Les (2020).

ways for finding justice in judgements that are not based on some universal criteria. The simplest illustration is the behaviour of the gods of Greek pantheon, whose reasons for action have no set universal criteria, yet seem to follow a set of rules throughout. One dichotomy to consider as an entry point is a distinction Lyotard makes between justice and piety. Piety here can stand for any metanarrative that tends to create criteria for truth. There can only be universal criteria for truth when we have a universal narrative within which that truth is understood. When the universal narratives become discredited and no longer function, there is no longer universal criteria for truth. Yet even without objective truth, we still want justice.

Another way to think about Lyotard's paganism is in opposition to *home*. Home is where the set, safe narratives exist. They do not change easily if at all. Home provides shelter and safety and in that sense it provides a stable narrative, if not a universal one. According to Lyotard, *Pagus* "was used to refer to the frontier region on the edge of towns" (Lyotard, 1992, p. 135). That's a place that is not entirely wild, but not entirely safe. "You do not expect to discover the truth there; but you do meet lots of entities who are liable to undergo metamorphoses, to tell lies, and to become jealous or angry: passable gods." (Lyotard, 1998, p. 136). What this metaphor tries to illustrate is that there are some frameworks that are not stable, that cannot be stable, because they do not have a stable grounding or some timeless reference point. In other words, they are not universal. To bring us back to the topic in simple terms: there is no universal criteria for marking the essays, but we still want to mark them justly.

A further point that needs to be grasped from Lyotard's paganism is the function that narratives perform when the supposed universality of criteria is gone. Continuing with the metaphor, Lyotard tells us how pagans came to terms with their gods - with the forces whose criteria is beyond anyone's understanding. Instead of appealing to the gods as universal arbiters of justice, pagans "came to terms with them by way of counter-plots, offerings, promises, and little marriage contracts that gave rise to complicitous ceremonies." (Lyotard, 1998, p. 136). What this means is that the narration is not one sided.

> "A pagan god, [...], is an effective narrator. You hear a story you are being told; it makes you laugh, cry or think, it inspires you to do something, to undertake a certain action, to put off making a decision, or to tell yourself a story. The narrator forces you into one or another of the narrative instances; he makes you a listener, an actor or a story-teller. That is where his superior strength lies; he manipulates you like a sorcerer; that is your weakness; you are dependent upon him; you have to get by with the stories he tells you and makes up for you" (Lyotard, 1998, p. 137).

The pagan god is just like the setter of the marking criteria - an effective narrator that forces you to participate in the narrative. But the criteria are not given in a universal form, they cannot be.[4] All "criteria" (reasons) in marking are given as narratives - subject competency, understanding, rationalisation are all subject to interpretation, especially for the best students. There is no universal judgement to be derived from these rubrics and only reasons for making a particular judgement in a particular case.

[4] Due to the irreducibility of the narrative interpretation.

From this pagan perspective that Lyotard gives us, we see that what may at times appear to us as judgements based on universal criteria is merely an effective narrative. Once we recognise that it is all narration, we get to reply. A reply here is not merely a reaction, but an elevation of a given narrative to a new language game. This may be best understood from the perspective of the one being judged - a student. The marking rubric does not provide them with a criteria on how to write the essay, merely a language game against which they have to write. The essay does not simply uncover some truth - it either participates, elevates or denies the prescribed language game. Lyotard says of the way pagans talked to their gods: “They talked in order to produce certain effects, not in order to profess the truth, to uncover an uncovering or to confess their guilt” (Lyotard, 1998, p. 136). However, with automated marking as a judge the production of narratives is severely limited.

Postmodern Condition:

If the writing of essays for the assignment is not merely to show the knowledge of some truth, but to produce certain narratives, one may ask the question of why we must judge at all. Two different (but not necessarily exclusive) narratives will be considered here. One from the perspective of performativity as a criterion, the other from innovation and experimentation as an end in itself.

In *The Postmodern Condition* Lyotard reflects on the direction that education takes via the influence of performativity as a criterion. The basic premise here is that when the performativity “of the supposed social system is taken as the criterion of relevance [...], higher education becomes a subsystem of the social system, and the same performativity criterion is applied” (Lyotard, 2004, p. 48). In other words, the criterion of efficiency becomes the leading language game. I suggest that this is the same framework that pushes the pursuit of automated marking. As Lyotard says: “The question now asked by the professional student, the State, or institutions of higher education is no longer “Is it true?” but “What use is it?” (Lyotard, 2004, p. 51). This may seem paradoxical at first - automated systems’ ‘use’ value and ‘truth’ value is not intrinsically separate. The truth value of an automated system will always be a set of criteria derived for efficiency. That is why Perelman is able to make an algorithm that fools automated marking - it exploits the efficiency-oriented nature of algorithmic judgements. The judgement of ‘is it true’ becomes prefaced on the judgement of its function (or use).

Beyond the simple issue of an “efficient exploitation of efficiency”[5] within algorithmic judgements, there is a serious issue of what performativity depends on in the final analysis.

The conflict identified by Lyotard is that “any experimentation in discourse, institutions, and values is regarded as having little or no operational value and is not given the slightest credence in the name of the seriousness of the system.” (Lyotard, 2004, p. 50). Meaning that a system based on performativity is reluctant to allow experimentation, even when experimentation may allow for better performativity in the long run. In the context of essays and marking, encouraging experimentation would produce better results in the long run, but

[5] Meaning that an algorithm's reliance on using the most efficient relations between concepts leaves it open to exploitation on that basis through other algorithms.

undermine the narrative of performativity at the same time. Ultimately this points to a broader issue of marking as a mode of evaluation of learning in performance driven society, but that's for another time.

Furthermore, "given equal competence [...], what extra performativity depends on in the final analysis is imagination" (Lyotard, 2004, p. 52). Imagination[6] allows one to make new moves, form new narratives and displace any set criteria as inadequate for a universal judgement. While such imagination is technically possible for an automated system to produce, without a set criteria for the evaluation, it is doomed to be unable to recognise the best performing essays.

Kant's judgements:

So far I have shown that the issue with automated marking is the impossibility of deducing a set of universal criteria. The reasons for marking (as well as writing) a particular way are always open to interpretation and new language games. Before I go into explaining the role language games play in marking further, I will briefly explore Kant's understanding of the act of judging. This is done in order to suggest that the judgement of an essay's mark is at least partly an aesthetic judgement driven by regulating Idea, which should never be confused with a concept. Such interpretation allows opening the discussion for the interplay of different language games.

Even in the first Critique, Kant shows the inadequacy of general logic providing the rules for judgement.

> "General logic contains, and can contain, no rules for judgement. For since general logic abstracts from all content of knowledge, the sole task that remains to it is to give an analytical exposition of the form of knowledge [as expressed] in concepts, in judgements, and in inferences, and so to obtain formal rules for all employment of understanding. If it sought to give general instructions how we are to subsume under these rules, that is, to distinguish whether something does or does not come under them, that could only be by means of another rule. This in turn, for the very reason that it is a rule, again demands guidance from judgement." (Kant, 2018, A133/B172).

Simply put, in order to interpret a given rule for judgement, we still have to judge how and whether the rule should be applied.

According to Kant in the third critique, judgements of beauty are based on a particular kind of pleasure. Namely, disinterested pleasure, meaning that the subject has no desire for the object, but finds it beautiful anyway. Furthermore, the reason why the judgement of a mark (especially a good one) can be considered as an aesthetic judgement is because, as established prior, marking does not follow any universal criteria, but relies instead on a multitude of reasons for interpretation. Marking judgement takes the same form as a judgement of beauty, in that it is subjective, but also necessary. The aesthetic judgement is not based on determinate concepts or rules. According to Kant, this type (aesthetic) of

[6] Imagination is read here in the Kantian sense of productive/reproductive synthesis.

judgement depends on “free play” of the faculties of imagination and understanding. As per Lyotard’s reading of Kant, the notion of a regulating Idea is always a reflective use of judgement.[7] That is, “a maximisation of concepts outside of any knowledge of reality” (Lyotard, 1985, p. 75*)*. The Idea “is not even able to give us contents for prescriptions, but just regulates our prescriptives, that is, guides us in knowing what is just and what is not just” (Lyotard, 1985, p. 77). To bring it to automation and put it in simple terms, it takes a reflective judgement to be able to mark essays that escape the imposed language game, something that automation cannot do as it would require the algorithm to form higher-order representations. The reflective judgement is guided by the Idea, which in the case of marking could be said to be the maximisation of the concept of justice.

I note that the point of this interpretation of Kant is not to suggest that marking is solely an aesthetic judgement. All that is needed for my argument is that evaluation of the best essays has the elements of aesthetic judgement as just described, because this shows the lack of universal criteria in at least some domains of the high end essay writing.

Language games:

The idea of language games shows us what education should be like. A language game is essentially a category or a framework with a set of rules. The rules must be followed for the particular language game to exist, but every diversion from a given set of rules can be considered as a new language game. Some language games are better than others, depending on how many different combinations of utterances they allow or how well they allow one to be understood. Lyotard sees language games as being a fight, in the sense of playing, where one tries to win. One does not have to play to win by necessity, as Lyotard says: “Great joy is had in the endless invention of turns of phrase, [...] but undoubtedly even this pleasure depends on a feeling of success won at the expense of an adversary - at least one adversary [...] the accepted language or connotation.” (Lyotard, 2004, p. 10).

Two brief elements are important to note here. First is that although language can be said to communicate information, it should not be reduced to simply that. Reducing language to its function as mode of communication risks privileging the system’s own interest and point of view.

The second element to note is that language games are agonistic, meaning that they are made against an “adversary”, they are made in order to “win”. This is important as it helps us recognise that every utterance pertaining to someone evokes a displacement which necessitates a response. The response can be of two kinds - a reaction or a reply. The difference between the two is that a reaction essentially remains within a given language game. A reply, on the other hand, elevates the communication to a new language game. As Lyotard put it: “Reacting means insulting someone who insults you. Replying means you triumph when someone insults you.” (Lyotard, 1998, p. 137).

Conclusion:

[7] Reflective judgement means going from a particular individual/concept to a more universal judgement.

The distinction between reaction and reply is important when we consider essays and their marking, especially automated marking. It is important, because an automated system will never be able to form a reply and is merely reactionary.

Now, when we put the pieces together, everything should make sense. The pagan view - one which denies the metanarratives as legitimate knowledge and encourages a multitude of language games - is a useful tool, because it shows that the criteria by which we judge are not universal.

The criterion of performativity, so prevalent in higher education today, is stifling imagination and creativity. This is problematic not just for learning, but also for judging. As the brief exposition of Kantian judgements shows, marking is an active, reflective endeavour guided by an Idea as horizon, not as a determinate concept.

With the addition of the language games as agonistic and participatory, we can see how hopelessly inadequate automated marking would be. Not only would it stifle creativity already under attack by the narrative of performativity, it would generate an appeal to a specific language game, defined by insufficient criteria that only appear as universal and that could only be reacted to, but never replied to. The writing of philosophical essays in the assignment form is not just an exercise of knowledge, but a chance to think and generate a reply to the language games set out by the course and the thinkers that came before.

Bibliography:

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

Kant, I. (2018). *Critique Of Judgement*. S.L.: A & D Publishing.

Lyotard, J.F. and Bennington, G. (2004). *The postmodern condition : a report on knowledge*. Minneapolis, Minn. Univ. Of Minnesota Press.

Lyotard, J.F. and Benjamin, A.E. (1998). *The Lyotard reader*. Oxford: Blackwell.

Lyotard, J.F. and Thébaud, J.L. (1985). *Just gaming*. Minneapolis: University Of Minnesota Press.

Perelman, L. (2012). Chapter 7. Construct Validity, Length, Score, and Time in Holistically Graded Writing Assessments: The Case against Automated Essay Scoring (AES). *International Advances in Writing Research: Cultures, Places, Measures*, [online] pp.121–132. doi:10.37514/per-b.2012.0452.2.07.

Perelman, L. (2020). "The BABEL Generator and E-Rater: 21st Century Writing Constructs and Automated Essay Scoring (AES)". *Journal of Writing Assessment*, 13(1). Retrieved from https://escholarship.org/uc/item/263565cq