“F**k the algorithm”?: What the world can learn from the UK’s A-level grading fiasco

The A-level grading fiasco in the UK led to public outrage over algorithmic bias. This is a well-established problem that data professionals have sought to address through making their algorithms more explainable. However, ***Dr Daan Kolkman*** *argues that the emergence of a “critical audience” in the A-level grading fiasco poses a model for a more effective means of countering bias and intellectual lock-in in the development of algorithms.*

Last week, hundreds of students in UK gathered in front of the Department for Education and chanted “f**k the algorithm”. Within days, their protests prompted officials to reverse course and throw out test scores that an algorithm had generated for students who never sat their exams due to the pandemic.

This incident has shone the media spotlight on the question of AI bias. However, previous cases of AI bias have already led to well-intentioned efforts by data scientists, statisticians, and machine learning experts to look beyond the technical and also consider the fairness, accountability, confidentiality, and transparency of their algorithms. What the A-level grading fiasco demonstrates is that this work may be misdirected. There is a key lesson to be learned from this algorithmic grading fiasco. A lesson that will only become more relevant as governments and organizations increasingly use automated systems to inform or make decisions: There can be no algorithmic accountability without a critical audience. By this, I mean that, unless it draws the attention of people who critically engage with it, technical and non-technical quality assurance of algorithms is a token gesture and will fail to have the desired effect.

The UK’s A-level grading fiasco

For those who haven’t followed the story – two weeks ago, thousands of students in England and Wales received their “A-level” exam grades. Instead of scoring actual exams, however, grades were determined by an algorithm. Almost 40% of students received grades lower than they had anticipated, sparking public outcry and legal action. Faced with protests, the UK government retracted the grades. Students will now receive grades based on their teacher’s estimate of what their grade would have been, had the exams gone forward as planned.

Date originally posted: 2020-08-26

Permalink: https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/

Blog homepage: https://blogs.lse.ac.uk/impactofsocialsciences/

How did the algorithm discriminate?

So what happened in the case of the A-level exam grading fiasco? Since COVID-19 hit the UK, student numbers for the 2020-2021 academic year were expected to drop by as much as 24%. This will cause a huge drop in tuition income for universities and lead to increased competition between them. In response, the UK government capped growth rates for the number of students at each university at 5% and introduced substantial financial penalties for any university that exceeded that limit.

At the same time, social distancing measures meant that the so-called “A-level” exams could not go ahead as planned. These exams play a pivotal part in the admissions procedure of UK academic institutions. Students around age 18 apply to universities prior to taking the exams and receive offers conditional on their exam scores. A difference of one grade can mean they won’t get into their university of choice.

In the absence of actual exams, Ofqual decided to estimate the A-level grades using an algorithm. Three inputs were used: 1. The historical grade distribution of schools from the three previous years (2017-2019); 2. The rank of each student within her own school for a particular subject, based on a teacher’s evaluation of their likely grade had the A-levels gone forward as planned (called the “Centre Assessed Grade” or CAG for short); 3. The previous exam results for a student per subject.

The algorithm looked at the historical grade distribution of a school and then decided a students’ grade on the basis of their ranking. For instance, if a student was halfway down the ranking list, then her grade would be roughly equal to what the person in the same ranking obtained in previous years. This approach was intended to correct for observed grade inflation in the CAGs (pdf), which explains why the algorithm’s grades were lower than the scores students expected. However, the use of historical data in algorithms is also a key component in latent algorithmic bias.

> The effects of many other algorithms are much less pronounced and may only be felt by small groups of people who are unlikely to find support in the media. This presents a key accountability challenge standing in the way of responsible use of algorithms, a challenge we need to solve.

Date originally posted: 2020-08-26

Permalink: https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/

Blog homepage: https://blogs.lse.ac.uk/impactofsocialsciences/

Several people identified issues with Ofqual’s algorithm from a technical report that was released by the UK government. Among other things, experts criticized the low accuracy of the algorithm and lack of uncertainty bounds for the resulting grades. Meanwhile, public outcry centred on the algorithm’s unfair results. For instance, if no one from your school has gotten the highest grade in the past three years, it’s extremely unlikely—if not impossible—for anyone from your school to attain that grade this year.

In addition, the algorithm puts more weight on the CAGs if there are fewer than 15 students in a particular subject at a particular school. That meant students at smaller schools were more likely to benefit from grade inflation than those at larger schools. This approach reinforces existing inequalities, as one analysis showed that the “proportion of A* and As awarded to independent (fee-paying) schools rose by 4.7 percentage points—more than double the rate for state comprehensive schools.”’

Anyone who has ever developed an algorithm will know George Box’s cautionary words by heart: “All models are wrong”. No algorithm is perfect and thus Ofqual should have anticipated the possibility of unfair or unjust outcomes to some students. However, when news about the A-level grades broke, there was no clear appeals procedure in place. Rather, the procedure was very complicated and students had to pay in order to appeal their grades. Like the model itself, the flaws in the appeals procedure were likely to disproportionately affect students from lower socio-economic backgrounds. The situation was further exacerbated since universities couldn’t take in more students due to the growth rate cap. Universities could not afford to be more flexible in their admissions even if they wanted to.

> Many of the cases of algorithms gone rogue that we know about could have been stopped by critical reflection earlier in the process. Such reflection however, is unlikely to come introspectively. Despite their best efforts, those developing algorithms will be prone to bias and intellectual lock-in.

Algorithmic accountability

I have focused my own research on algorithmic decision making and accountability in the public sector and have witnessed firsthand the amount of effort that goes into developing, testing, and implementing algorithms. Data professionals go to great lengths to ensure the validity and robustness of their models. Great work is underway to develop, amongst other things, new tools to explain even the deepest neural nests.

However, what stood out to me about the A-level exam grading fiasco was the massive public outcry and protest. The media, educators, the students, and their parents put enormous pressure on the UK government to reconsider the automatically determined grades. They formed **“a critical audience”** that collectively scrutinized the algorithm, the underlying data, and the wanting procedures for redress.

This is not to argue against efforts to make algorithms more explainable or develop quality assurance processes for ever more complex models. My argument here is that these tools, explanations, and processes run the risk of remaining empty signifiers if no critical engagement emerges from their installment. Such critical engagement emerged in the case of A-level grading algorithm, because the impacts of that algorithm were directly felt by thousands of students.

> critical engagement emerged in the case of A-level grading algorithm, because the impacts of that algorithm were directly felt by thousands of students

The role of a critical audience

Date originally posted: 2020-08-26

Permalink: https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/

Blog homepage: https://blogs.lse.ac.uk/impactofsocialsciences/

Not all algorithms used in government are problematic, but few are as visible as Ofqual’s grading algorithm. The effects of many other algorithms are much less pronounced and may only be felt by small groups of people who are unlikely to find support in the media. This presents a key accountability challenge standing in the way of responsible use of algorithms, a challenge we need to solve.

Yes, it’s certainly worth scrutinizing the data and methodologies behind automated systems. Yes, tools for explainable algorithms are a welcome addition to our toolkit. Yes, we need to think about the fairness, accountability, and transparency of algorithms. Our efforts should not end there. We need to think carefully about how we can create critical audiences for the millions of algorithms that impact our daily live.

> Without a critical audience that opposes algorithms and points out their shortcomings, we will keep hearing about the occasional incident with automated decision making, but never learn of the majority of algorithms which screw-ups never see the light of day.

Many of the cases of algorithms gone rogue that we know about, could have been stopped by critical reflection earlier in the process. Such reflection however, is unlikely to come introspectively. Despite their best efforts, those developing algorithms will be prone to bias and intellectual lock-in. It is precisely for this reason that quality assurance guidelines don’t work; we cannot grade our own work, or that of people we know really well.

Without a critical audience that opposes algorithms and points out their shortcomings, we will keep hearing about the occasional incident with automated decision making, but never learn of the majority of algorithms which screw-ups never see the light of day. The fostering of such critical audiences meets with its own share of issues. Do we instate an algorithm regulator? How would it deal with propriety systems, the high costs of reviews, how would it find out where algorithms are used?

These are difficult questions that we need to talk about. The victory over the A-level algorithm is no happy ending. UK universities will likely struggle with increased competition, owing to the cap. And if public officials everywhere don’t start to think more seriously about fostering critical audiences for algorithms , I’m afraid we’ll hear, “F**k the algorithm!” far more often in the future.

Note: This article gives the views of the authors, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below

Photo credit: Markus Spiske on Unsplash

Date originally posted: 2020-08-26

Permalink: https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/

Blog homepage: https://blogs.lse.ac.uk/impactofsocialsciences/