The UK School Algorithm Debacle: Five Lessons for Corporate AI Programs

August 25, 2020

The widespread criticism, and partial abandonment, of the algorithm that was used to evaluate UK students serves as useful reminder that corporate AI programs carry significant regulatory and reputational risks, and that careful planning, testing and governance are needed throughout the process to mitigate those risks.

BACKGROUND

In March, due to the pandemic, UK authorities canceled the exams that students usually take at the end of primary and secondary school. These exams are critical in determining which secondary school or university students will attend. Indeed, most students had offers from secondary schools and universities that were conditional upon them achieving a certain result in the exams that were canceled. So a new system had to be devised to award graded degrees to hundreds of thousands of students.

The system that was devised required schools to give each student (1) a predicted grade (referred to as "centre-assessed grades", or "CAGs") and (2) a rank from best to worst (in their subject). That information was then sent to the government regulator ("Ofqual") for a process of "moderation". Ofqual had been tasked by the UK government to develop, in its words, "a system for awarding calculated grades, which maintained standards and ensured that grades were awarded broadly in line with previous years," which was viewed as a mandate to keep grade inflation in check.

Ofqual developed an algorithm through which all CAGs and school rankings were fed. The algorithm compared CAGs with the performance of students at each individual school over the past three years. The algorithm then "moderated" (*i.e.*, lowered, maintained or increased) each CAG to ensure consistency between individual students' CAGs and historical results at their schools, which then resulted in the grade awarded to the individual student. For reasons of statistical accuracy, the smaller the relevant subject-class, the less "interventionist" the algorithm, such that for classes with five students or fewer, the algorithm simply accepted CAGs.

Through this process, approximately 40% of students had their grades lowered as compared with their CAGs. This led to a nationwide outcry, largely because the methodology appeared to have exacerbated social inequality by generally favoring students at private schools at the expense of their state-sector peers. First, private schools historically perform better than state schools. Second, private school classes are significantly smaller. Because the algorithm broadly ensured that students could not significantly out-perform the average of their predecessors, a top student in a historically low-achieving school could not really score better than the average student at her/his rank over the past three years. Conversely, students at historically high-achieving schools generally did not underperform their historical peers.—

This resulted in dozens of individual stories about high-achieving students in deprived areas marked down by the algorithm, which had capped their performance at their schools' previous highs, and thereby prevented them from meeting conditional offers from prestigious universities. Eventually, the government backtracked and amended the methodology such that students could rely on either their CAGs or their moderated grades, whichever was higher.

FIVE LESSONS FOR CORPORATE AI AND ALGORITHMS DECISION-MAKING

Problems Can Arise Even When the Model Is Behaving as Expected

Most risk assessments for automated decisions focus on model drift or other unexpected behavior. But the UK school algorithm worked exactly as intended. It kept grade inflation in check and maintained a distribution within predictable ranges. The problem was in the design—a failure to appreciate the expectation that decisions affecting individuals must be primarily based on individualized factors.

Automated Decision Making Risks Are Not Limited to AI

Most of the regulatory focus on algorithms concerns the development and adoption of AI and machine learning, but the UK schools algorithm is neither. Rather, it is a complex, but static, model that did not learn or develop on its own, and therefore serves as a reminder that any automated decision-making will likely be subject to intense scrutiny when it can significantly impact the lives of individuals (e.g., hiring, lending, promotions, salary, etc.).

Good Faith Will Not Be an Adequate Defense

The main criticisms of the algorithm are not based on any bad faith. The pandemic presented the UK government with a serious problem that needed a quick solution that could be applied on a grand scale, and had to provide certainty and fairness, while not

resulting in an unacceptable burden on the UK school and university system. The technocrats who designed the system honestly believed they had achieved those goals. But none of that shielded them from withering criticism when the results were seen to be unfair and biased in favor of privileged students.

The Importance of Limiting Outliers

What seemed really to galvanize the protests were particular cases that appeared indefensible: high-achievers from historically underperforming schools whose scores had been significantly lowered. While these outliers were clearly predicable from the outset, the modelers failed to appreciate their likely significance. Data scientists tend to have a utilitarian view of model outcomes—the larger the percentage of good decisions, the better. But the tolerance for outliers in human decision-making cannot serve as the baseline for algorithms because of the perceived interconnectedness of automated decisions. If a teacher gives an exceptional student an unfair grade on an essay, that does nothing to undermine students' grades in a different class, let alone the grades in another school in a different part of the country. But any unfair or absurd algorithmic decision can undermine faith in every decision made by the system. Indeed, one New York state court decided to strike down a teacher assessment algorithm for its failure to account for the effect of outliers on the model's ability to grade teacher performance.

It is for this reason that AI assessments increasingly include stress testing, black swan modelling and the insertion of guardrails to limit the number of outliers that can undermine the entire project. Companies should also consider testing for bias in groups that may be underrepresented in the underlying datasets, as the model's decisions towards these outlier populations may be particularly skewed or severe.

The Need for a Human Review or Appeals

Public dissatisfaction with the algorithm's results was compounded by an opaque and confusing appeals process. For instance, students were told they could use their mock-exam results as a basis for an appeal. However, not all schools organized mock exams, and there was little consistency in approach among the schools that did. There was also confusion as to who should or could appeal (the student individually or the school on her/his behalf) and who should bear the costs of the appeals. Social acceptance for algorithmic and AI decision-making often requires clear and readily accessible ways to appeal to a human.

CONCLUSION

The failure of the UK government to anticipate the problems that its school algorithm would face resulted in enormous stress for hundreds of thousands of families, teachers,

education administrators and politicians. It led to precisely the situation the algorithm was initially designed to prevent: significant grade inflation, with 38% of students taking English A-levels being awarded the top A or A* grades, compared to 25% last year, with adverse knock-on effects on universities as well as future school graduates. It also seems to have undermined tolerance of algorithms more generally. *The Guardian* has reported a trend of UK public bodies abandoning recently introduced algorithmic systems for visa applications, benefit claims and other social welfare issues—largely based on concerns about bias, transparency and fairness. The lessons for corporations is that risk management, including testing and governance, should be carefully considered for any model that significantly affects individuals.

* * *

Please do not hesitate to contact us with any questions.

NEW YORK

Avi Gesser
agesser@debevoise.com

Anna R. Gressel
argressel@debevoise.com

LONDON

Robin Lööf
rloof@debevoise.com