SciencesPo
ÉCOLE D'AFFAIRES PUBLIQUES

Étude de cas

RESPONSIBLE USE OF ALGORITHMS IN THE PUBLIC SECTOR

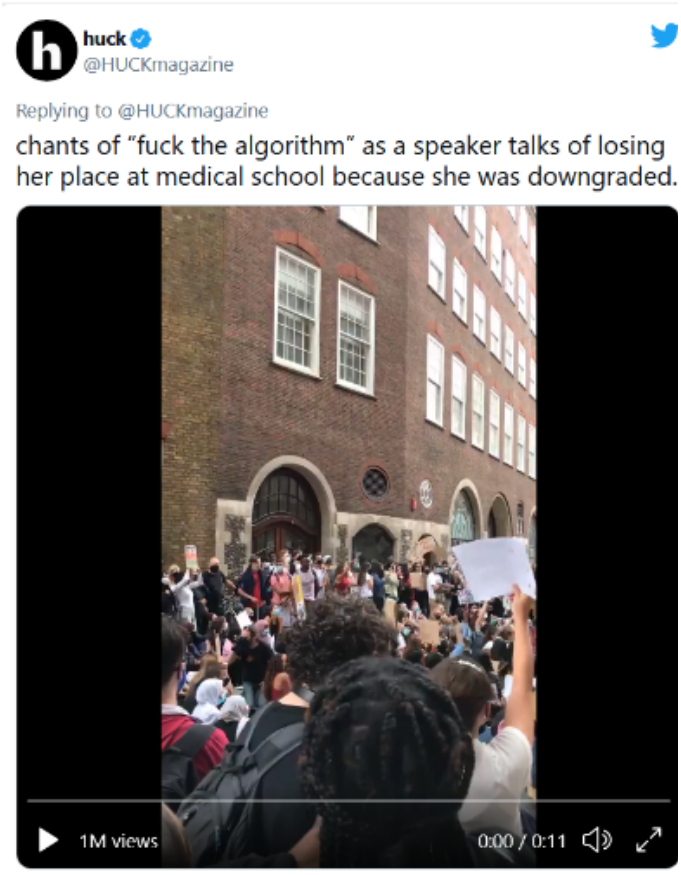
CASE STUDY OF THE ENGLISH 2020 EXAM RESULT ALGORITHM

Soizic Pénicaud

Introduction.“F**k the algorithm”: understanding England’s 2020 A-level controversy

A.Implementing responsible algorithmic decision-making systems: drawing lessons from a counterexample

On August 16 2020, hundreds of English[1] students went out on the street to protest against their government’s process to assign A-level grades during the covid-19 pandemic. More specifically, they targeted the system at the heart of the process: a predictive algorithm which determined the grades of over 280,000 students[2]. This resulted in unprecedented protest chants specifically targeting an “algorithm”:



Smoke, B. [@bencsmoke] for Huck [@huckmagazine]. (2020, 16 August). [Tweet, video]. Twitter. https://twitter.com/HUCKmagazine/status/1294985562106015750

[1] This case refers to “England” and not to the United Kingdom as Scotland, Wales and Northern Ireland have different ways of assessing students and this algorithm is specific to England.
[2] BBC News. (2020, 14 August), *A-levels: Labour call for government U-turn over 'exams fiasco'.* https://www.bbc.com/news/education-53776938

More and more national and local governments have implemented digital algorithms[3] in the past few years, in fields ranging from social benefits to education to healthcare[4]. Their hopeful intent: using technology to make public service more efficient and fairer. This take often overlooks the issues that arise with the implementation of digital decision-making systems in government, such as biased decisions that pass off as objective and lack of transparency and oversight[5].

However, despite these issues, automated-decision making systems (ADMS) (see insert) shouldn't be automatically discarded from the public sector. This bears the question: **how can governments responsibly design and implement automated-decision making systems?**

The A-level ADMS controversy (deemed a "fiasco" by some sources[6]) spans March - September 2020, although the aftermath continued to be felt and debated throughout 2020-2021. It illustrates a large number of issues that can arise when governments implement ADMS, all the more in times of crises. **This case is an opportunity to observe the real-life effects of these issues and think about how to mitigate them.**

What is an automated decision-making system?

Nota: there is still a lot of discussion in the field about which word to use to describe the systems at stake. Here, we will purposely stay away from the term "artificial intelligence" and use the term "automated-decision making (ADM) systems" to emphasize that the tools are the basis for governmental decisions and to take the focus away from the technologies used, which can range from simple spreadsheets to more complex machine learning algorithms* (although they entail specific problems).*

Words followed by an asterisk (*) are defined in the annex at the end of the case study.

To learn more about "artificial intelligence" and "machine learning", have a look at the following additional resources, depending on how you learn best (video, text, comic strip):

- Jist Studios, BBC Ideas. (2019, 19 September). "What exactly is an algorithm?" [Video]. https://www.bbc.co.uk/ideas/videos/what-exactly-is-an-algorithm/p07nw8ny
- Leufer, D. (2020). *Myth: the term AI has a clear meaning*. AI Myths. https://www.aimyths.org/the-term-ai-has-a-clear-meaning
- Julia Stoyanovich and Falaah Arif Khan. "What is AI?". We are AI Comics, Vol 1 (2021). https://dataresponsibly.github.io/we-are-ai/comics/vol1_en.pdf

[3] Referred to in this case study as "automated decision-making systems", or ADM systems.
[4] For examples in Europe, see AlgorithmWatch. (2020). *Automating Society Report 2020.* *https://automatingsociety.algorithmwatch.org/*
[5] See O'Neil, C. (2016)," Bomb parts: What is a model?" and "Shell shocked: My journey of disillusionment". In *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.*
[6] See, among others, Kolkman, D. (2020, 15 August). *"F**ck the algorithm? What the world can learn from the UK A-level grading algorithm fiasco".* LSE Impact Blog. https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/ ; Murkett, C. (2021, 25 February). "Prepare for the next A-level fiasco". *The Spectator*. https://www.spectator.co.uk/article/prepare-for-the-next-a-level-fiasco

(p.4 of the comic)

B.Public policy & technology: how to quickly adapt the university admissions process during a time of crisis?

England's higher education admissions system[7]

To understand the stake of the ADM system studied here, it is important to understand how higher education admissions work in England.

Universities accept different types of secondary education graduation diplomas (another one being the International Baccalaureate), with the A-levels being the most common. Typically, a student will have to take at least 3 A-level subjects to enter university. Students who take the A-levels are graded on a scale of A* (the best grade) to E (the worst grade).

In their final year of secondary education, students, typically aged 17-18, who want to apply to university receive predicted grades from their teachers. For example, a student studying

[7] UCAS. (2021). *Filling in your UCAS undergraduate application*. UCAS. https://www.ucas.com/undergraduate/applying-university/filling-your-ucas-undergraduate-application

Physics, Maths and English may receive a predicted grade of A* in Physics, A in Maths and B in English. They will refer to these grades as A*AB.

Students apply to a number of universities through the UCAS system, where they submit their transcripts, a personal statement as well as the predicted grades from their teachers.

Universities then send out offers to students before the results of the exams. These offers are usually conditional, which means they are tied to the applicant achieving a certain result in their final exams, based on their application and university standards. For example, the aforementioned student may be conditionally accepted to the (fictional) University of Leeds-City if they achieve AAB in their A-levels.

Students and universities receive the results of their A-levels in August, and university starts in October. Final exams therefore play a crucial part in higher education admissions in the UK.

Due to the covid-19 crisis, it was impossible to organize exams for A-levels. On March 18th 2020, Secretary of State for Education Gavin Williamson simultaneously announced the closing of schools and the cancellation of exams. *The Guardian* journalist Richard Adams relates that Williamson told the Commons:

> *"I can confirm that we will not go ahead with assessments or exams, and that we will not be publishing performance tables for this academic year. We will work with a sector and have to ensure children get the qualifications that they need.*
> *My department is working closely with local authorities, representatives of early years schools and head teachers, regional schools, commissioners and bodies such as Ofqual about how to deliver this change as effectively as possible".*

Source : Adams, R. (2020, 18 March). All schools to close from Friday; GCSE and A-level exams cancelled – UK Covid-19, as it happened. *The Guardian*. https://www.theguardian.com/politics/live/2020/mar/18/uk-coronavirus-live-boris-johnson-pmqs-cbi-urges-government-pay-businesses-directly-saying-350bn-loangrant-package-not-enough?page=with:block-5e7261318f088d7575595edc#block-5e7261318f088d7575595edc

Government and the non-ministerial department Office of Qualifications and Examinations Regulation, commonly known as Ofqual, were faced with the challenge of finding a solution to ensure a smooth transition for students between secondary school and higher education, in a very short timespan.

In late March 2020, Gavin Williamson, the Secretary of State for Education, asked head of Ofqual, Sally Collier, to "ensure, as far as is possible, that qualification standards are maintained

and the distribution of grades follows a similar profile to that in previous years"[8]. On 31 March, he issued a ministerial direction under the Children and Learning Act 2009[9].

Ofqual resorted to an ADM system to predict student's A-level grades. In a guidance document on April 3rd, they gave information on some of the data they intended to use:

> *"To make sure that grades are as fair as possible across schools and colleges, exam boards will put all centre assessment grades through a process of standardisation using a model being developed with Ofqual. We will consult on the principles of our model shortly, but we expect it will look at evidence such as the expected national outcomes of this year's students, the prior attainment of students at each school and college (at cohort, not individual level), and the results of the school or college in recent years. It will not change the rank order of students within each centre; nor will it assume that the distribution of grades in each subject or centre should be the same. The process will also recognise the past performance of schools and colleges. However, if grading judgments in some schools and colleges appear to be more severe or generous than others, exam boards will adjust the grades of some or all of those students upwards or downwards accordingly".*

Source: Ofqual (2020, April 3rd). *How GCSEs, AS & A levels will be awarded in summer 2020.* Gov.uk. https://www.gov.uk/government/news/how-gcses-as-a-levels-will-be-awarded-in-summer-2020

It is worth noting that resorting to an ADM was in and of itself a choice. Other countries or private baccalaureate systems also resorted to algorithms, such as the International Baccalaureate[10] or Scotland[11]. On the other hand, countries like France just decided to forgo exams and only rely on continuous assessment[12].

[8] Adams, R.; Elgot, J.; Stewart, H.; Proctor, K. (2020 19 August). "Ofqual ignored exams warning a month ago amid ministers' pressure". *The Guardian*. https://www.theguardian.com/politics/2020/aug/19/ofqual-was-warned-a-month-ago-that-exams-algorithm-was-volatile

[9] Direction under S 129(6) of the Apprenticeships, Skills, Children and Learning Act 2009. (2020 21 March). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/877611/Letter_from_Secretary_of_State_for_Education_to_Sally_Collier.pdf

[10] Simonite, T. (2020, 10 July). *Meet the Secret Algorithm That's Keeping Students Out of College.* Wired.com. https://www.wired.com/story/algorithm-set-students-grades-altered-futures/

[11] Esson, G. (2020, 4 August). *Scotland's results 2020: How grades were worked out for Scottish pupils.* BBC News. https://www.bbc.com/news/uk-scotland-53580888

[12] Cojean, T. (2020, 5 April). *Annulation des épreuves du bac : un événement historique rarissime.* L'Etudiant. https://www.letudiant.fr/bac/annulation-des-epreuves-du-bac-un-evenement-historique-rarissime.html

Although a public consultation was conducted[13] from April 15-29th 2020, Ofqual remained opaque about the parameters of the algorithms until the results.

However, it did release a useful 319-page interim report after the grades were assigned, on Thursday August 13th 2020[14].

Among other things, this report describes the inner workings of the model.

C.A standardized ADM system to ensure a "fair" evaluation of students

After deciding to implement an ADMS, Ofqual had several choices to make with regards to what type of model*, input data* and standardization method to use. All these choices led to the development of the "Direct Centre Performance model", described in the interim report.

Ofqual's premise was that teachers tend to overestimate their students' grades. To avoid this, the algorithm didn't take into account the teacher's predicted grades (referred to as "center-assessed grades", or "CAGs").

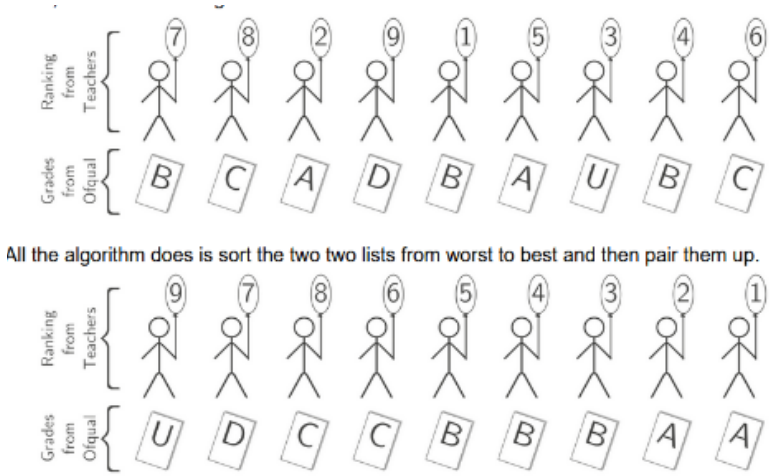
> *"We considered many different options, but it was apparent that the best judges of the relative ability of students in a school or college were the teachers who had been preparing these students for their exams, tracking their progress relative to target grades, and, in the case of A level students applying to higher education, providing estimated grades.*
>
> *We therefore asked teachers to provide, for each student for each subject they were entered for, a centre assessment grade (CAG) which represented the grade that student would have been most likely to achieve if teaching and learning had continued and students had taken their exams as planned.*
> *We also asked teachers to provide a rank order of students for each grade for each subject. There were several reasons for this. First, we know from research evidence that people are better at making relative judgements than absolute judgements and that teachers' judgements tend to be more accurate when they are ranking students rather than estimating their future attainment. The research literature suggests that, in estimating the grades students are likely to achieve, teachers tend to be optimistic*

[13] Ofqual. (2020). *Consultation: Exceptional arrangements for exam grading and assessment in 2020.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879627/Exceptional_arrangements_for_exam_grading_and_assessment_in_2020.pdf

[14] Ofqual. (2020). *Research and Analysis: Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909368/6656-1_Awarding_GCSE__AS__A_level__advanced_extension_awards_and_extended_project_qualifications_in_summer_2020_-_interim_report.pdf

(although not in all cases). That is not surprising, teachers want to do the best for their students, and the analysis we carried out immediately after CAGs were submitted bears this out."[15]

To summarize: Ofqual asked the teachers to rank the students in their class (list 1 in Haine's schema below), and combined this ranking with the school grading's history from 2017-2019 (list 2 in the schema below) and a standardization method[16] to determine individual students' grades (the algorithm).

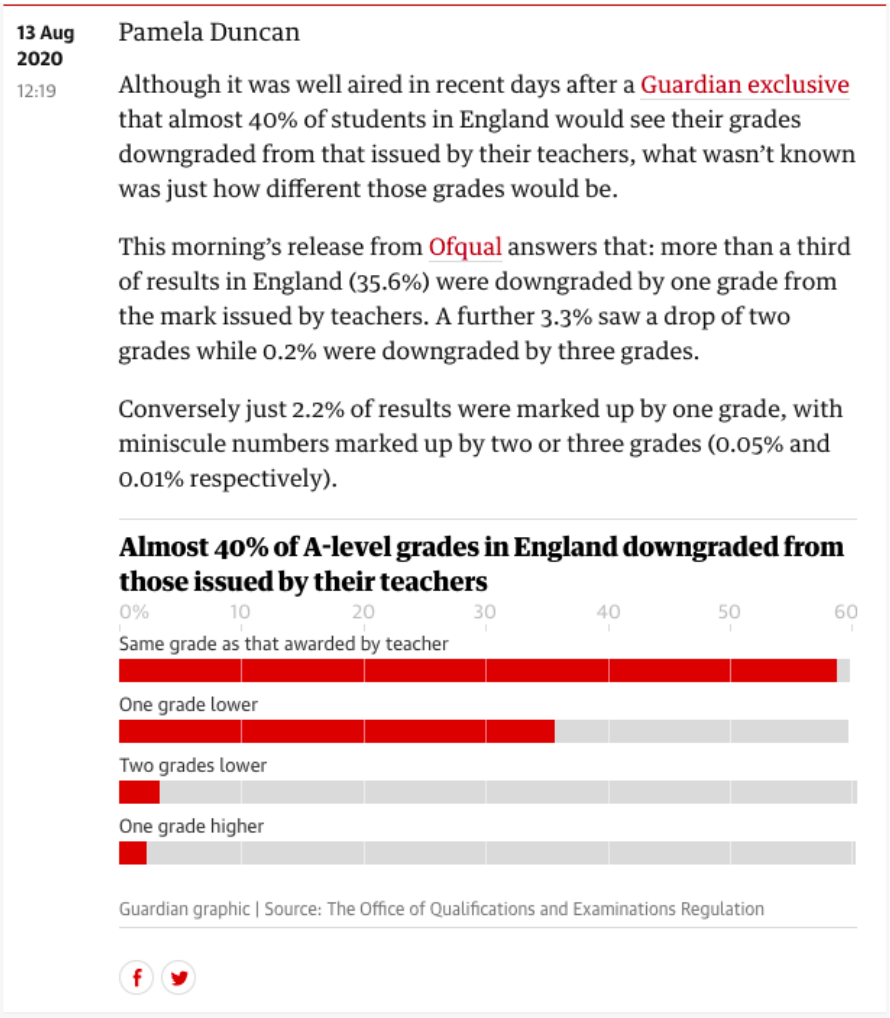


Source: Haines, T. *A-Levels: The Model is not the Student*. Tom SF Haines' website. http://thaines.com/post/alevels2020

On the day of the results, many students were shocked to receive lower grades than expected. This was especially widespread among students from lower-income backgrounds.

[15] Source: Ofqual (2020, 13 August). *Executive summary: Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909035/6656-2_-_Executive_summary.pdf. pp. 1-2

[16] */!\ If you're getting scared at this point, don't panic! You don't need to know more on a technical level to understand this case. If you're curious to learn more, see* Bennett, S. H. (2020, 20 August). *3. Reliance on rank order in On A Levels, Ofqual and Algorithms.* Sophie Bennett's blog. https://www.sophieheloisebennett.com/posts/a-levels-2020/



13 Aug 2020
12:19

Pamela Duncan

Although it was well aired in recent days after a Guardian exclusive that almost 40% of students in England would see their grades downgraded from that issued by their teachers, what wasn't known was just how different those grades would be.

This morning's release from Ofqual answers that: more than a third of results in England (35.6%) were downgraded by one grade from the mark issued by teachers. A further 3.3% saw a drop of two grades while 0.2% were downgraded by three grades.

Conversely just 2.2% of results were marked up by one grade, with miniscule numbers marked up by two or three grades (0.05% and 0.01% respectively).

Almost 40% of A-level grades in England downgraded from those issued by their teachers

Guardian graphic | Source: The Office of Qualifications and Examinations Regulation

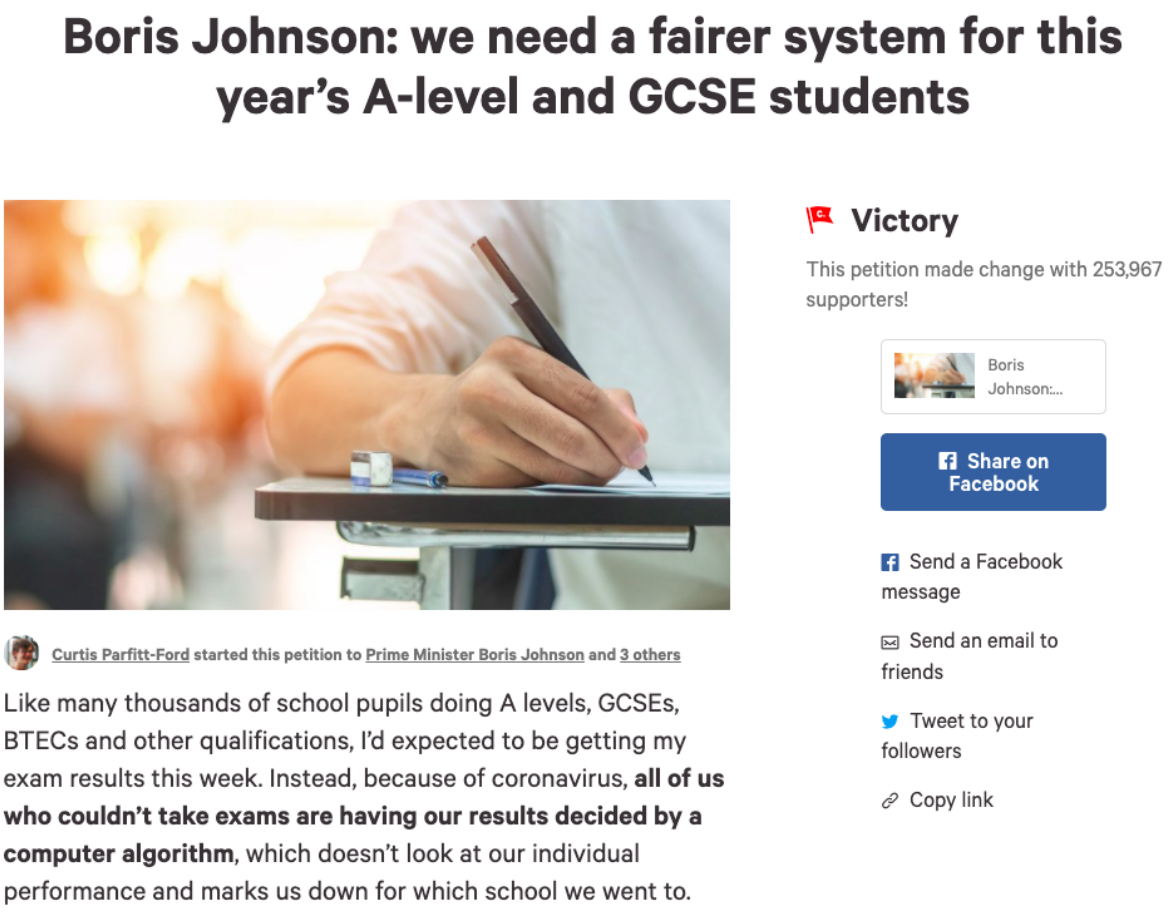
Source: Duncan, P. (2020, 13 August). *A-level results day 2020 live: 39.1% of pupils' grades in England downgraded - as it happened*. The Guardian. https://www.theguardian.com/education/live/2020/aug/13/a-level-results-day-2020-live-students-teachers-government-ucas-mock-exams-triple-lock-nick-gibb?page=with:block-5f351f6b8f08899e2e66d6b5#block-5f351f6b8f08899e2e66d6b5

D.Civil society strikes back: mobilization against the government

Ofqual's interim report[17] enabled specialists (data scientists, academics, journalists and digital rights NGOs) to look into the inner workings of the algorithms and to point out several flaws.

A student named Curtis Parfitt-Ford started a petition which received over 250 000 signatures. Around the same time, large protests sparked around the country.

[17] See Ofqual (2020). *Research and Analysis...*



Screenshot from the petition:
https://www.change.org/p/boris-johnson-boris-johnson-we-need-a-fairer-system-for-this-year-s-a-level-and-gcse-students-alevelresults?lang=en-US

On August 12th 2020, legal firm Foxglove announced through a press release that it had sent out a legal letter on behalf of Curtis Parfitt-Ford, demanding government to *"Grade the student, not the school: threat of legal action unless Ofqual fix unfair A-level and GCSE grading algorithm"*[18].

After being initially reluctant, the government backtracked on August 17th 2021:

> *"Gavin Williamson has announced that all A-level and GCSE results will be based on teachers' predictions as he apologised to students, parents and schools.*

[18] Foxglove. (2020, 12 August). *Press release: UK: Legal action threatened over algorithm used to grade teenagers' exams.* Statewatch.org. https://www.statewatch.org/news/2020/august/uk-legal-action-threatened-over-algorithm-used-to-grade-teenagers-exams/

During a press conference, the education secretary said he was "sorry for the distress this has caused" after a dramatic retreat by the government over the system for awarding grades."

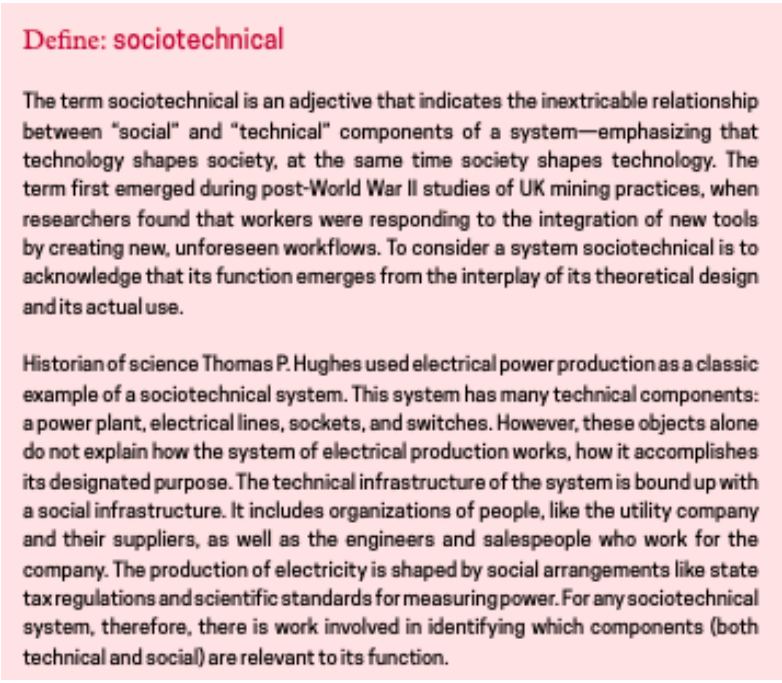
Source: Bennett, R. and Steven Swinford. (2020, 17 August) *Gavin Williamson apologises as he backs down on A-level and GCSE results*. The Times. https://www.thetimes.co.uk/article/gavin-williamson-apologises-as-he-backs-down-on-a-level-and-gcse-results-05zp3jvl3

Why did this ADM system undergo such strong backlash? What can we learn from it?

E.Studying this automated-decision making system under a sociotechnical lens

This case study draws from academic and grey literature (newspaper articles, blog posts, etc.), and comparisons with other cases. We will use a combination of analytical frameworks: political science, science and technology studies and critical data science and experimental methods inspired by the field of design.

One of the central concepts we will use to study the case is the science and technology notion of "sociotechnical system":



Define: sociotechnical

The term sociotechnical is an adjective that indicates the inextricable relationship between "social" and "technical" components of a system—emphasizing that technology shapes society, at the same time society shapes technology. The term first emerged during post-World War II studies of UK mining practices, when researchers found that workers were responding to the integration of new tools by creating new, unforeseen workflows. To consider a system sociotechnical is to acknowledge that its function emerges from the interplay of its theoretical design and its actual use.

Historian of science Thomas P. Hughes used electrical power production as a classic example of a sociotechnical system. This system has many technical components: a power plant, electrical lines, sockets, and switches. However, these objects alone do not explain how the system of electrical production works, how it accomplishes its designated purpose. The technical infrastructure of the system is bound up with a social infrastructure. It includes organizations of people, like the utility company and their suppliers, as well as the engineers and salespeople who work for the company. The production of electricity is shaped by social arrangements like state tax regulations and scientific standards for measuring power. For any sociotechnical system, therefore, there is work involved in identifying which components (both technical and social) are relevant to its function.

Source: Elish, M.C. and Elizabeth Anne Watkins, Data & Society. (2020). *Repairing Innovation: A Study of Integrating AI in Clinical Care.* https://datasociety.net/wp-content/uploads/2020/09/Repairing-Innovation-DataSociety-20200930-1.pdf. p.11

With this lens in mind, we will go more in depth into the issues posed by the ADM system (part I), the way the public policy was implemented and the politics behind the seemingly-only-technical system (part II), and the manner in which the government responded to the strong civil society backlash (part III).

I.From design to implementation to redress: the many problems of the A-level ADM system

Let us break down the criticisms made towards the system, first with regards to data and model design (A), then to technical quality assurance (B), then to accountability of policy makers and transparency of the process (C) and finally to lack of adequate appeals and redress processes (D).

A. From input data to model development: was the ADM system fair?

As the Office for Statistics Regulation Authority points out: "*it was always going to be extremely difficult for a model-based approach to grades to command public confidence. The task of awarding grades in the absence of examinations was very difficult. There were numerous challenges that the qualification regulators and awarding organisations had to overcome."*[19]

What were those challenges? How could (some) of them have been overcome?

1. Historical bias: students from lower-income backgrounds are worse off

Digging deeper into the criticisms made towards the biases in the ADM system outlined in the introduction, we'll see that the government and Ofqual seemingly omitted to take into account societal harm at different stages of the "machine learning life cycle"*[20], despite the Equality impact assessment conducted[21].

[19] Office for Statistics Regulation Authority. Ensuring statistical models command public confidence: Learning lessons from the approach to developing models for awarding grades in the UK in 2020. Executive summary. https://osr.statisticsauthority.gov.uk/publication/ensuring-statistical-models-command-public-confidence/

[20] Suresh, H. and John Guttag. (2020). *"A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle", Proceedings from the 2020 conference on Fairness, Accountability and Transparency of Machine Learning*, Barcelona.

[21] Ofqual's Strategy Risk and Research Directorate (2020). *Research and Analysis: Equality impact assessment: literature review*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879605/Equality_impact_assessment_literature_review_15_April_2020.pdf

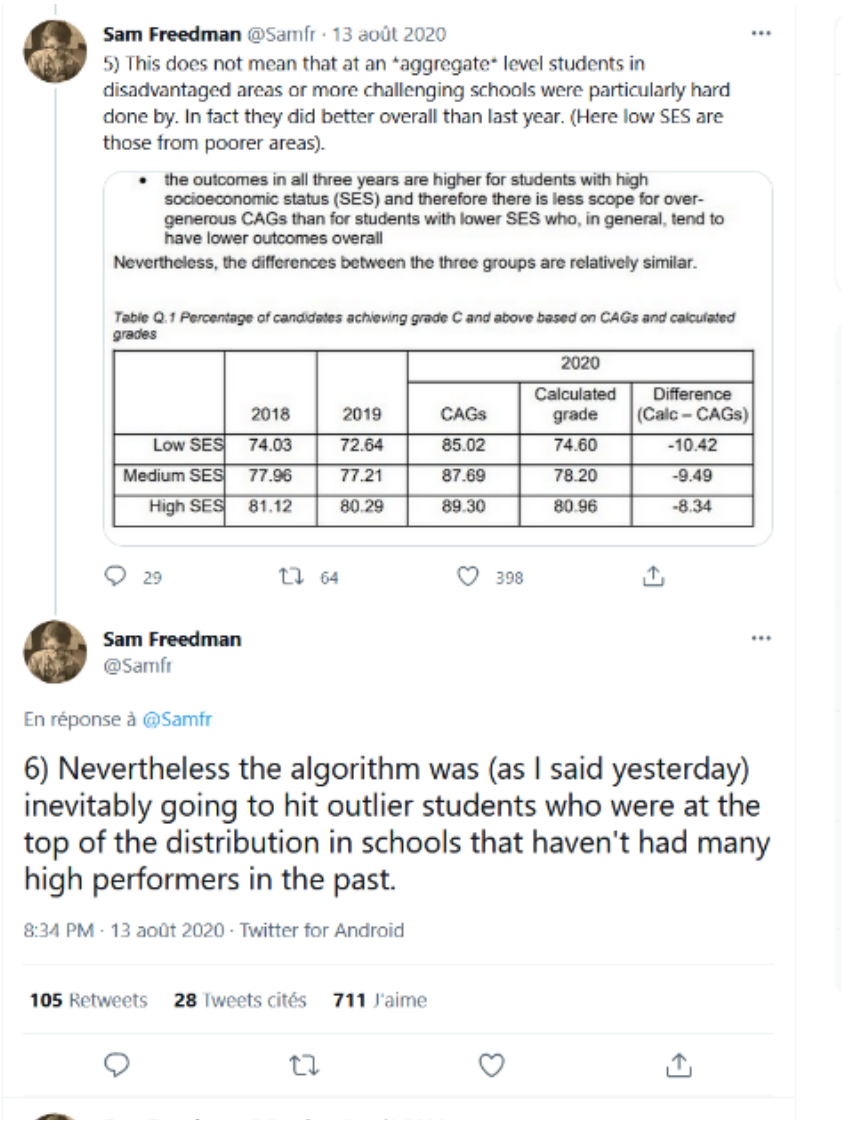
The first type of bias we can notice with the A-level ADMS is historical bias[22]: as the system cannot assess the students' actual performance in 2020, it has to use a proxy*. Here, it relies on historical data from schools, assuming students can only be as good as their predecessors (and not better).

Factoring in the schools' history into the student's final grades is at the crux of the problem. As Dr. Ben Green frames it: *"Although most people talk about machine learning's ability to predict the future, what it really does is predict the past."*[23]

The choice to use historical school data to predict individual grades had a high impact on **outliers* :**

[22] See Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems, cited in Julia Stoyanovich and Falaah Arif Khan. "All about that Bias". We are AI Comics, Vol 4 (2021). https://dataresponsibly.github.io/we-are-ai/comics/vol4_en.pdf

[23] Green, B. Z. (2019). *The Smart Enough City: Putting Technology in Its Place to Reclaim Our Urban Future, Initiales*. MIT Press.



Sam Freedman @Samfr · 13 août 2020

5) This does not mean that at an *aggregate* level students in disadvantaged areas or more challenging schools were particularly hard done by. In fact they did better overall than last year. (Here low SES are those from poorer areas).

- the outcomes in all three years are higher for students with high socioeconomic status (SES) and therefore there is less scope for over-generous CAGs than for students with lower SES who, in general, tend to have lower outcomes overall

Nevertheless, the differences between the three groups are relatively similar.

Table Q.1 Percentage of candidates achieving grade C and above based on CAGs and calculated grades

	2018	2019	2020		
			CAGs	Calculated grade	Difference (Calc – CAGs)
Low SES	74.03	72.64	85.02	74.60	-10.42
Medium SES	77.96	77.21	87.69	78.20	-9.49
High SES	81.12	80.29	89.30	80.96	-8.34

29 64 398

Sam Freedman
@Samfr

En réponse à @Samfr

6) Nevertheless the algorithm was (as I said yesterday) inevitably going to hit outlier students who were at the top of the distribution in schools that haven't had many high performers in the past.

8:34 PM · 13 août 2020 · Twitter for Android

105 Retweets **28** Tweets cités **711** J'aime

Freedman, S. [@Samfr]. (2020, August 13). “Nevertheless the algorithm was (as I said yesterday) inevitably going to hit outlier students who were at the top of…” [Tweet]. Twitter. https://twitter.com/Samfr/status/1293979304179769346

What this means in practice is that the model* developed by Ofqual is that an extraordinary student on their way to get an A* but attending a school whose best students had historically never gotten past an A had no chance of receiving an A*.

This happened often for students attending historically lesser good schools, often in unprivileged areas.

Data scientist Dr. Sophie Bennett summarizes it as such:

> *"In some schools, performance may look very similar year on year, making it not totally unreasonable to try and predict the new cohort's grade distribution from historic data. However, in other schools, performance may vary more widely year on year. In general, disadvantaged state schools are more likely to see larger variation in grades from one year to the next. However, Ofqual's algorithm completely ignores the variability in performance year on year. For instance, it's possible for performance to vary in a school such that a student achieves an A* this year despite no students having achieved an A between 2017-2019. However, the algorithm cannot account for events like this. This is yet another way that state schools in more deprived areas are disadvantaged by the model. (and 2. The algorithm does not seem to properly account for differences in the value-added across schools.)"*

Bennett, S. H. (2020, 20 August). *On A Levels, Ofqual and Algorithms.* Sophie Bennett's blog. https://www.sophieheloisebennett.com/posts/a-levels-2020/

2.Unequal implementation/technical bias: students from better-off schools are better-off

As seen in the introduction, Ofqual resorted to a standardization model* to ensure homogeneity of grades across at a national level, with equality in mind.

However, for statistical reasons, smaller classes were exempted from this standardization formula using past school results.

Channel 4's fact-checking section"[24] delved deeper into how the standardization model chosen by Ofqual put better-off schools at an advantage, without intending to do so. Ofqual responded to FactCheck:

> *"Our standardisation model, for which we have published full details, does not distinguish between different types of centres [schools], and therefore contains no bias, either in favour or against, a particular centre type (...). One of the factors for the increases in higher grades for some centres will be if they have smaller cohorts because teachers' predictions are given more weight in these circumstances (...) Centre assessment grades [teacher's predicted grades] are the most reliable source of evidence for small cohorts and is the fairest approach available. While use of the model is the fairest possible way to calculate grades for larger cohorts, it would not have been feasible to apply it to very low entries which would include special schools, pupil referral units and tutorial centres and small subject cohorts within larger centres."*

The reason smaller cohorts were exempted from the standardization formula is because, to work out statistically reliable patterns over time, one has to have a greater number of "data points" (here, students). Therefore, classes under 15 students relied on teachers' predictions.

As FactCheck then points out:

[24] Lee, G. (2020, 14 August). Did England exam system favour private schools?. *4 News FactCheck* .https://www.channel4.com/news/factcheck/factcheck-did-england-exam-system-favour-private-schools

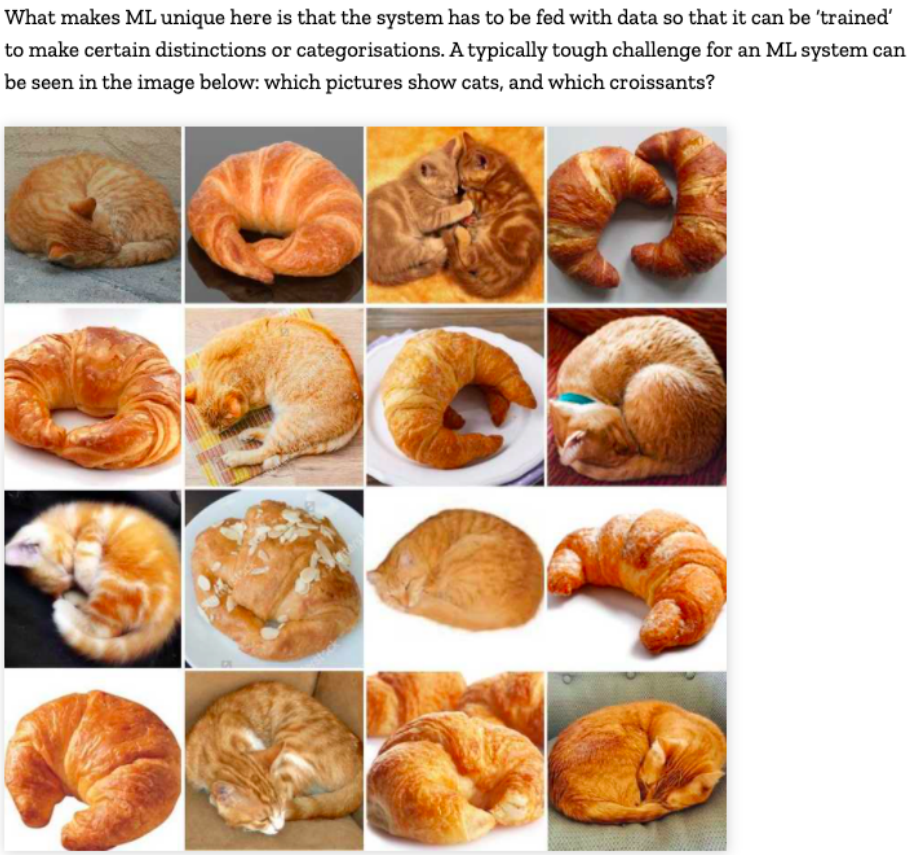
> "*On average, teachers across all types of school were more generous to students than the grades produced by the regulator's statistical model. So any student who's exempted from moderation – or has its effects diluted – is likely to do better (...) In practice, this means that if you're one of a very small number of students taking a particular subject at your school, you're likely to be awarded the more generous grade your teacher estimated for you (or something close to it).*"

And it highlights that *"small cohorts can exist at any type of school, but they seem to be more prevalent at fee-paying ones."*[25]

B. Lack of quality assurance

Quality assurance and technical accuracy are pillars of public trust in public sector algorithms.

Despite popular belief, "artificial intelligence" can be highly inaccurate, as illustrated in the following image by Dr. Daniel Leufer[26].



[25] Ibid.
[26] Leufer, D. (2020). *Myth: AI can solve any problem.* AI Myths. https://www.aimyths.org/ai-can-solve-any-problem

This means that special attention has to be paid to the quality of the technical tool itself, through testing it and evaluating its efficacy and its effects (the same way toy manufacturers test the robustness of the toys by hitting them as many times as they can in a controlled environment).

To measure the accuracy of an algorithmic model, one typically compares the results obtained by the algorithm and the results obtained without the algorithm (through a controlled pool of data that the algorithm is not trained on). Here, the accuracy was tested by comparing 2020 grades and 2019 grades.

On that topic, British think tank Ada Lovelace Institute released a blogpost outlining several opportunities for improvement, one of which being ensuring technical accuracy.

> *"One of the key parts of building trust in algorithmic systems is ensuring they provide accurate results. Ofqual's model seemingly failed here, though perhaps unavoidably.*
>
> *Ofqual stated that their model had about 60% predictive accuracy on average across A-level subjects, meaning that they expect 40% of grades would have been different had exams been sat, after testing the model on 2019 data.This seems unacceptable low accuracy.*
>
> *But evaluation shows this level of accuracy was broadly comparable to the probability of examiners awarding marks to students' exam papers that result in the same grade awarded by a senior examiner's marking [8][27]. So the model is no less variable for most students than traditional examination marking, and interrogating the level of accuracy exposes the underlying uncertainty in assessing students at a single point in time."*[28]

The Office for Statistics Regulation's observations concur and put emphasis on ensuring quality assurance for outliers (as we've seen in part A):

> *"In the exam case, there were clear examples of good quality assurance of both input and output data. For input data, centres were provided with detailed guidance on the data they should supply. For output data, the regulators undertook a wide range of analysis, largely at an aggregate level. There was limited human review of outputs of the models at an individual level prior to results day. Instead, the appeal process was expected to address any issues. There was media focus on cases where a student's grade was significantly different from the teacher prediction. In our view, these concerns were predictable and, whilst we recognise the constraints in this scenario, such cases should be explored as part of quality assurance."*[29]

---

[27] Nota: this information can be found on pp.80-81 of the interim report.
[28] Jones, E. and Safak, C. (2020, 18 August). *Can algorithms ever make the grade?.* Ada Lovelace Institute's blog. https://www.adalovelaceinstitute.org/blog/can-algorithms-ever-make-the-grade/
[29] Office for Statistical Regulation Authority (2020), prev. cited.

Commentators have also pointed out that properly testing the accuracy of the algorithm was not possible in this case:

> *"The normal way to test a predictive algorithm is to run it against the previous year's data: this was not possible as the teacher rank order was not collected in previous years. Instead, tests used the rank order that had emerged from the 2019 final results."*
>
> Harkness, T. (2020, 18 August). *How Ofqual failed the algorithm test*. Unherd. https://unherd.com/2020/08/how-ofqual-failed-the-algorithm-test/

There can be a form of magical belief around algorithms, thinking that because they resort to mathematics, they are right. However, although this case study emphasizes the importance of humans and politics in technical systems, the accuracy of these systems should not be underestimated, as the systems are usually way less efficient than they claim to be.

C. Transparency theater: lack of scrutiny and public engagement

After the design and testing of the model, Ofqual and the government were criticized for failing to implement the process properly, under the guise of participation and transparency.

1.Lack of transparency for external auditors

During the development of the model, the Royal Statistics Society offered to help Ofqual, but Ofqual conditioned this help with strong opacity. The Ada Lovelace Institute's report summarizes it as such:

> *"Another troubling issue raised throughout the development of the grading model has been the absence of independent, external scrutiny. The Royal Statistical Society's identification of concerns over the composition of the 'technical advisory group', consisting mainly of 'government employees or current or former employees of the qualification regulators',[30] and their subsequent offer to provide independent Fellows, was met with the condition of a strict, five-year non-disclosure agreement.*
>
> *As the RSS has indicated in their written submission to the Education Select Committee call for evidence on the impact of COVID-19 on education and children's services inquiry, 'without a stronger procedural basis to ensure statistical rigour, and greater transparency about the issues that Ofqual is examining, it cannot be clear that the statistical methodology will be beyond question".[31]*

[30] Quoting Ashby, D. and Witherspoon, S. (2020) 'Letter to Ed Humpherson, Director General for Regulation, Office for Statistics Regulation'. Available at: https://rss.org.uk/RSS/media/News-and-publications/News/2020/14-08-2020-Letter-Deborah-Ashby-Sharon-Witherspoon-to-OSR.pdf

[31] Ada Lovelace Institute (2020), previously cited.

2.“Engagement theater”

The process was not completely opaque and devoid of engagement. Indeed, a public consultation was held, which received almost 13 000 responses during the making of the algorithm, including schools, students, teachers, and unions[32]. The release of the interim report, published the day grades were released is also laudable: often, algorithms are obfuscated and never made public.

The problem is that after the consultation phase, Ofqual returned to an opaque way of developing the model until results day.

This illustrates an issue qualified as “engagement theater” by some in other contexts. For example, Bianca Wylie highlighted this in the controversy around the development of SideWalks Lab in Toronto, where she criticized the means of consultation put in place[33]. She explained at the time:

> *“Right now, residents don’t have a fighting chance of exerting their opinion on the smart city through public consultation, because these questions haven’t been answered, and the issues haven’t been explained or taught.”*

By keeping outside people in the dark, Ofqual may have missed a chance to, firstly, build public trust around its system and, secondly, receive useful insights *before* the algorithm was put in place.

D. A faulty appeal & redress process

The appeal procedure was only released when the students were given the grades. This defeats the purpose of an appeal process, which is supposed to help achieve an actual result (here: contest a decision to provoke change, for example to be accepted in their top-choice university).

As the Ada Lovelace Institute puts it:

> *“The systems of redress were inadequate: appeals should have been able to take place by schools in the weeks before grades were released, allowing the algorithmically assigned grades to be adjusted before they had a material impact. They also*

[32] Ofqual. (2020). Analysis of Consultation of Responses: Exceptional arrangements for exam grading and assessment in 2020 - Consultation on specified general qualifications – GCSEs, AS, A levels, Extended Project Qualifications and the Advanced Extension Award. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/886555/Analysis_of_consultation_responses_21MAY2020.pdf. p.7

[33] Wylie, B. (2018, 13 August). *Searching for the Smart City's Democratic Future*. Centre for International Governance Innovation. https://www.cigionline.org/articles/searching-smart-citys-democratic-future/

exacerbated inequalities, placing an additional burden on those unwilling to accept an unfair decision, and favouring those with the means to support an appeal."[34]

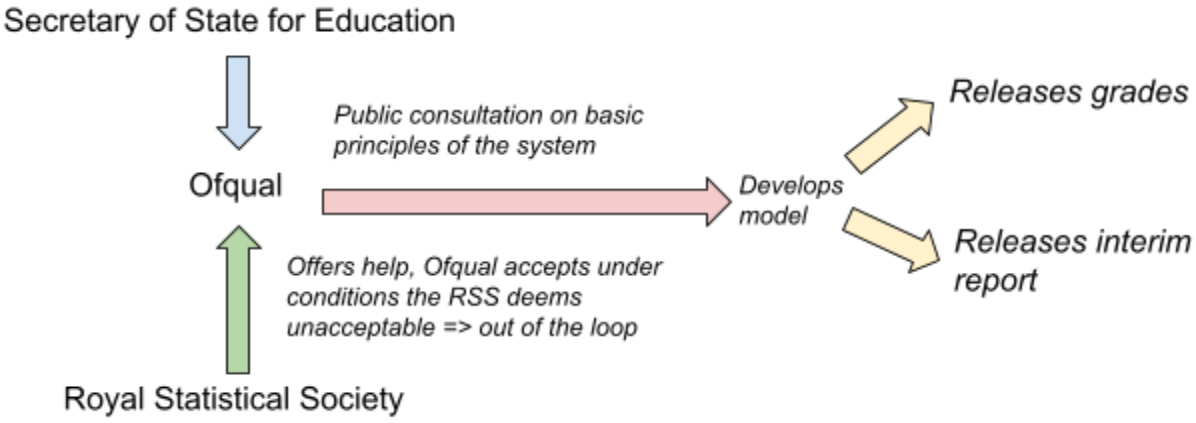
Other, more participatory solutions, could have been considered, such as the one by Dr. Sophie Bennett:

> *"Given the accuracy of the model, it should have been little surprise that many students would seek to appeal.Yet there was no appeal process in place when results were announced. Where the algorithm-estimated grades differed substantially from centre assessed grades, Ofqual could have also allowed teachers to pre-appeal on behalf of their students."*[35]

II."Black boxed politics": the choices behind the tech tools

As explained previously, the design and implementation of the A-level algorithm was heavily criticized, to the point of government backtracking.

How could this happen? What other choices could have been made? What does it reveal about the things policy makers should think about when they implement technology?



Actor mapping of the design & implementation of the process
Source: Soizic Pénicaud

[34] Jones, E. and Safak, C. (2020, 18 August). *Can algorithms ever make the grade?.* Ada Lovelace Institute's blog. https://www.adalovelaceinstitute.org/blog/can-algorithms-ever-make-the-grade/

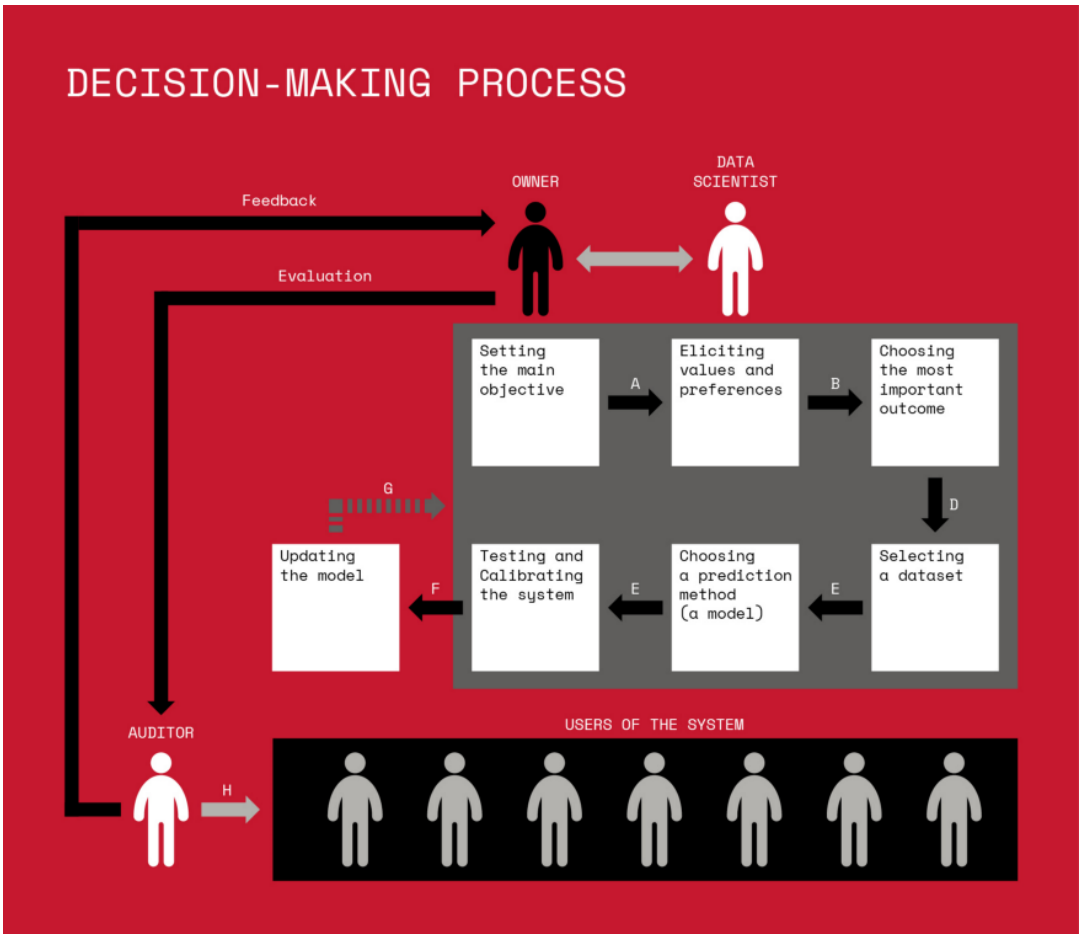
[35] Bennett, S. H. (2020, 20 August). *On A Levels, Ofqual and Algorithms.* Sophie Bennett's blog. https://www.sophieheloisebennett.com/posts/a-levels-2020/

A.The public policy choice: preventing grade inflation and ensuring fair distribution of grades across schools

One of the myths of algorithms is that they are unbiased and can do everything. However, one always has to optimize for something in algorithms, and choices are present at every step.

As Szymielewicz, Foryciarz and Leufer show, every tech tool development starts with a policy choice ("setting the main objective" in the schema below). In the case at hand, it was about preventing grade inflation and ensuring fair distribution across schools. As Secretary of State for Education put it: *"our goal has always been to protect the trust that public rightly has in education qualification*s"[36].

In order to do that, Ofqual developed a model using specific values and preferences, and choosing the main outcome.



[36] Weale, S. and Stewart, H. (2020, 17 August). *A-level and GCSE results in England to be based on teacher assessments in U-turn.* The Guardian. https://www.theguardian.com/education/2020/aug/17/a-levels-gcse-results-england-based-teacher-assessments-government-u-turn

Source: Szymielewicz, K.; Foryciarz A.; Leufer, D. (2020 January 17). *Black-Boxed Politics: Opacity is a Choice in AI Systems*. Medium [Blog].
https://medium.com/@szymielewicz/black-boxed-politics-cebc0d5a54ad

1.Preventing grade inflation

The first thing to remember is that Ofqual's model didn't take into account teacher's predicted grades, but a ranking they asked teachers to do. Raising the question as to why, Dr. Sophie Bennett found the following answer in pp. 13-21 of Ofqual's interim report:

- *"Differences in how strict centres were in generating grade predictions could lead to unfairness if too much weight was given to the centre assessed grades (page 20)*
- *Using centre assessed grades would cause significant grade inflation from last year to this year*
- *Psychological studies demonstrate that we are more accurate at making relative judgements (i.e. about rank) than we are at making absolute judgements*."[37]

In other words, according to the report, teachers tend to be too generous with their students, and relying on their predictions posed the risk of having too many students receive good grades (thus disrupting the university system).

2.Ensuring fair distribution across schools

Ofqual explains its choice to predict grade distribution for each school, rather than per student, in this way (as synthesized by Dr. Sophie Bennett based on Ofqual's interim report):

> *"The rationale for predicting **grade distributions for each school, rather than individual grades** is explained on pages 34-36.*
>
> *In this section, Ofqual describe three potential approaches to allocating grades:*
> - ***micro-level standardisation:** grade estimates are made for individual students based on the characteristics and performance of these individual students.*
> - ***meso-level standardisation:** grade estimates are made at a centre/school-level. This is the approach Ofqual settled on.*
> - ***macro-level standardisation:** apply a simple transformation to the centre assessed grades to standardise them, that is applied at a national level (i.e. the same transformation is applied to all centres/students in the same way).*
>
> *Ultimately, the report concludes that the macro-level standardisation too heavily on the centre assessed grades being accurate and fair across schools, and also cannot*

[37] Bennett, S. H. (2020, 20 August). *Ofqual and Algorithms.* Sophie Bennett's blog. https://www.sophieheloisebennett.com/posts/a-levels-2020/

satisfactorily account for differences in value-added across schools (...). Likewise, Ofqual rule out the micro-level standardisation, concluding that prior attainment is not a strong enough predictor of final grade, and that this approach could end up disrupting the rank orders provided by teachers."[38]

As shown above, the priority for the government was to ensure meso-level standardization as to prevent grade inflation while still taking into account some information provided by the teachers.

B.When reframing the policy at stake leads to not building an algorithm

Let us now look at some commentaries challenging this policy decision.

Journalist Karen Hao claims that the wrong policy choice was made in the first place:

> *But the root of the problem runs deeper than bad data or poor algorithmic design. The more fundamental errors were made before Ofqual even chose to pursue an algorithm. At bottom, the regulator lost sight of the ultimate goal: to help students transition into university during anxiety-ridden times. In this unprecedented situation, the exam system should have been completely rethought.*
>
> *"There was just a spectacular failure of imagination," says Hye Jung Han, a researcher at Human Rights Watch in the US, who focuses on children's rights and technology. "They just didn't question the very premise of so many of their processes even when they should have."*
>
> *At a basic level, Ofqual faced two potential objectives after exams were canceled. The first was to avoid grade inflation and standardize the scores; the second was to assess students as accurately as possible in a way useful for university admissions. Under a directive from the secretary of state, it prioritized the first goal. "I think really that's the moment that was the problem," says Hannah Fry, a senior lecturer at University College London and author of Hello World: How to Be Human in the Age of the Machine. "They were optimizing for the wrong thing. Then it basically doesn't matter what the algorithm is—it was never going to be perfect."*
>
> Source: Hao, K. (2020, 20 August). *The UK exam debacle reminds us that algorithms can't fix broken systems.* MIT Technology Review. https://www.technologyreview.com/2020/08/20/1007502/uk-exam-algorithm-cant-fix-broken-system/

[38] Bennett, S. H. (2020, 20 August). *Ofqual and Algorithms.* Sophie Bennett's blog. https://www.sophieheloisebennett.com/posts/a-levels-2020/

C.The right policy choice, but merely an extremely difficult task?

In their report, the Office of Statistics Regulation had a different opinion[39]:

> "*In our view, the teams within the qualification regulators and awarding organisations worked with integrity to try to develop the best method in the time available to them. In each country there were aspects of the model development that were done well, and aspects where a different choice may have led to a different outcome. However, none of the models were able to command public confidence and there was widespread public dissatisfaction of how the grades had been calculated and the impact on students' lives. (...)*
>
> *Our main conclusion is that achieving public confidence in statistical models is not just about the technical design of the model – taking the right decisions and actions with regards to transparency, communication and understanding public acceptability throughout the end to end process is just as important.*
>
> *We also conclude that guidance and support for public bodies developing models should be improved. Government has a central role to play in ensuring that models developed by public bodies command public confidence. This includes directing the development of guidance and support, ensuring that the rights of individuals are fully recognised and that accountabilities are clear.*"

Its official position is that it was possible to implement an algorithm, but that Ofqual's was unsatisfactory in its implementation.

However, by the institution's own admission, this was subject to a lot of internal discussions. During an online event organized by the Ada Lovelace Institute in May 2021, moderator of the panel Andrew Strait relayed a question from the audience: "was building a model inherently flawed?". Ed Humpherson, the Director General for Regulation of the Office for Statistics Regulation, explained:

> *"As we were doing this work, debate raged internally, (...) the debate was: "were the four bodies across the UK set an impossible task or merely an extremely difficult task?" (...) Our report lands in the latter without at all downplaying the enormous difficulties the organizations faced."*

[39] Office for Statistics Regulation Authority. Ensuring statistical models command public confidence: Learning lessons from the approach to developing models for awarding grades in the UK in 2020. Executive summary.
https://osr.statisticsauthority.gov.uk/publication/ensuring-statistical-models-command-public-confidence/

Interestingly, Gail Rankin, Head of Edinburgh Office & Systemic Review Programme Lead, underlined that it was not the Office's job to question the policy in place.[40]

Let us now delve into the time period after grades were distributed.

III. The aftermath: ADM systems as objects of political discontent

A. Rage against the algorithm: the opposition, experts & non-experts

At the same time as students were receiving unsatisfactory grades, Ofqual's interim report enabled researchers to understand how it worked.

On August 17, following public outcry, the government backtracked and announced they were *"scrapping the standardization model*".[41]

[40] Ada Lovelace Institute. (2021, 13 May). *Building public confidence in data-driven systems Findings of the Office for Statistics Regulation review into the 2020 exam results algorithm, and why public confidence in data-driven systems matters*. [Webinar]. https://www.adalovelaceinstitute.org/event/building-public-confidence-data-driven-systems/

[41] Weale, S. and Stewart, H. (2020, 17 August). *A-level and GCSE results in England to be based on teacher assessments in U-turn.* The Guardian. https://www.theguardian.com/education/2020/aug/17/a-levels-gcse-results-england-based-teacher-assessments-government-u-turn

A-level and GCSE results in England to be based on teacher assessments in U-turn

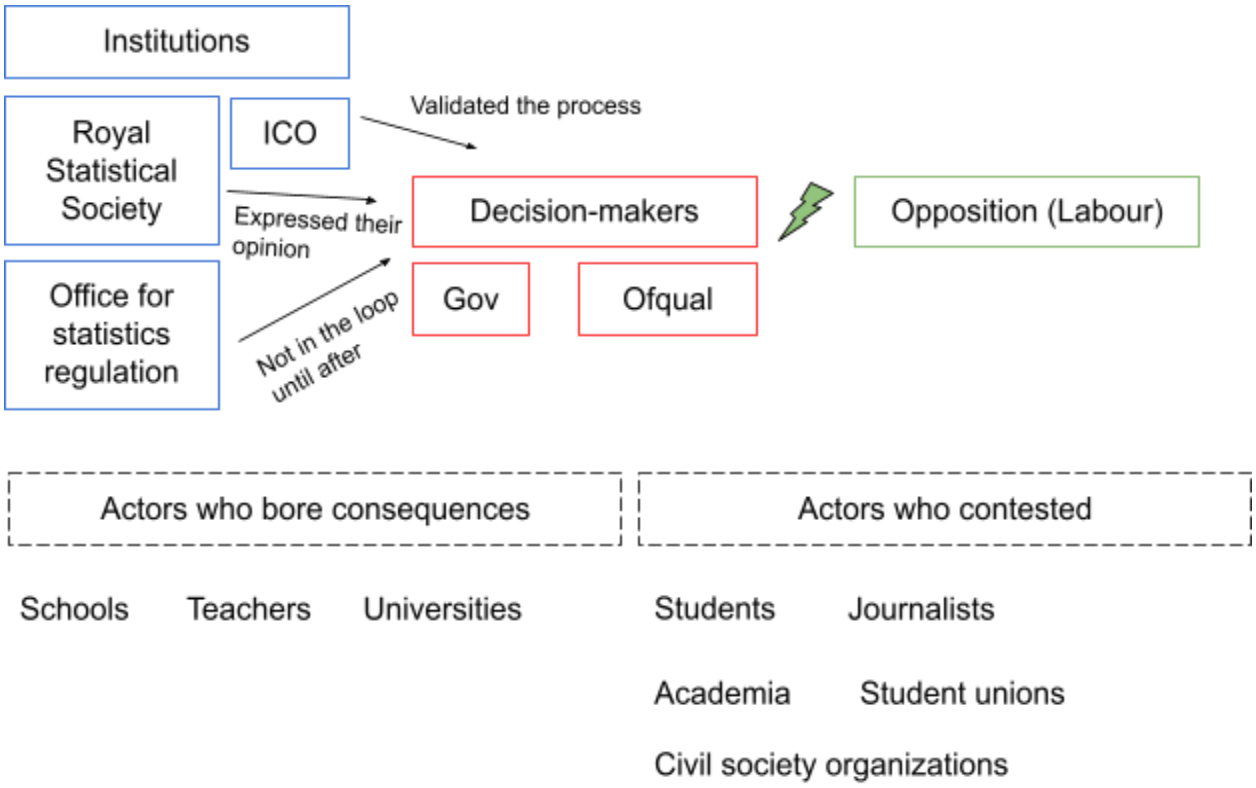
Williamson and Ofqual apologise, scrapping standardisation model after outcry

▲ A level students celebrate outside the Department for Education in London after it was confirmed that candidates in England will be given grades estimated by their teachers, rather than by an algorithm. Photograph: Victoria Jones/PA

Screenshot from Weale & Stewart (2020)

What led to such a sudden and drastic U-turn?

In August 17, several new players were already in the game:



Actor mapping of the period after grades were released.
Source : Soizic Pénicaud

Although the ICO, the UK's data protection authority, deemed that the system developed by Ofqual was not an "*automated-decision system*"[42], other actors contested the algorithm on different grounds. As early as August 14, Labour blamed the government for its "*fatally flawed results system*" and asked the government to go back to teacher-assessed grades[43].

B.On the government's side: the many stages of managing the backlash

As mentioned in the introduction, the government did not back down right away. On Thursday August 13 (the day of the results), Prime minister Boris Johnson started by denying all claims of unfairness.

[42] ICO. (2020, August 14). Statement in response to exam results. https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2020/08/statement-in-response-to-exam-results/
[43] BBC News. (2020, August 14), *A-levels: Labour call for government U-turn over 'exams fiasco'.* https://www.bbc.com/news/education-53776938

Boris Johnson insists A-level results are 'robust' as he shrugs off protests from teachers and pupils

Prime minister denies problems will hurt 2020 cohort, arguing students are snapping up the university places they are after

Rob Merrick Deputy Political Editor | @Rob_Merrick | Thursday 13 August 2020 17:20 | comments

Merrick, R. (2020, 13 August). *Boris Johnson insists A-level results are 'robust' as he shrugs off protests from teachers and pupils.* The Independent. https://www.independent.co.uk/news/uk/politics/boris-johnson-level-results-protests-teachers-pupils-gavin-williamson-a9669131.html

However, after the petition, the protests and the general backlash, the government was forced to backtrack on August 17. The story doesn't end there: on 25 August 2020, Sally Collier who oversaw the development of Williamson's algorithm calculation resigned from the post of chief regulator of Ofqual[44].

Something to be noted is that the issues caused by the ADM system didn't only impact students: after the government reverted to relying on teachers-assessed-grades, a lot of universities had to bear the brunt of this decision as they had already allocated all their slots to students on August 13. Edge cases such as independent candidates were also left at a disadvantage.

What's interesting is that, in a few days, Boris Johnson turned the narrative around. Addressing a class of pupils, he blamed the "mutant algorithm" that had put England and their upper-classmates in this position, saying: "*You couldn't sit your exams, which you yearned to do, your grades were almost derailed by a mutant algorithm.*"

[44] Richardson, H. (2020, 25 August). *Ofqual chief Sally Collier steps down after exams chaos*. BBC News. https://www.bbc.com/news/education-53909487

10 Downing Street. (2020, 28 August) *WATCH LIVE: PM Boris Johnson addresses school children in England (26/8/2020)*. [YouTube Video]. https://youtu.be/Q5BHMh7hyDY?t=668

Once again, we see that the tool is blamed for the policy at stake. The algorithm is demonized for something policy makers decided and for decisions whose consequences are irreversible.

Looking at this case may be disheartening. What can we learn from it? What does it tell us about the role of algorithms in the public sector and under which conditions they should be implemented?

C.Thinking critically to improve things: towards a "smart enough government"

As we've seen in many of the sources cited throughout the case study, independent commentators, think tanks, researchers and legal firms have challenged the myths around artificial intelligence, algorithms and automated-decision making as being a panacea. That being said, what principles should be put in place to ensure these technologies are used to improve things?

In 2018, researcher and former CTO in the public sector Ben Green wrote a book called the *Smart Enough City*. He argued that the general vision of the smart city was wrong and pursued the wrong goals. Instead, through different real-life examples of how data science could serve the public sector, he tried to define what the "smart enough city" would be like, *i.e.* a city that uses technology to pursue just and fair policies.

If we take this concept and apply it to the government, **what would the smart enough government look like? Are there examples of automated-decision making systems that are thought of and applied in a just, fair, transparent and accountable way?** What could we learn from them too?

The consequences of the 2020 A-level ADMS are still being felt by some, and the pandemic is not over. 2021 will be the opportunity to see if England and Ofqual have been able to learn from their mishaps and improved the way grades were distributed this year.

Annex: definitions of some technical terms

Algorithm: a process that applies specific rules to information entered at the beginning and produces a result at the end (like a baking recipe). Algorithms can be rule-based (the rules are defined and encoded by humans in a script, often digital) or data-driven (they learn the rules based on past examples). Data-driven algorithms are also called machine-learning algorithms.

Automated-decision making system: any system where an algorithm is involved. The system can be fully automated (*i.e.* the algorithm gives out the final result) or partially automated (*i.e.* an algorithm is involved at some point of the process, but the final result or decision is left to a human).

Artificial intelligence: currently popular word that encompasses a lot of different things. Most of the time, it is used to refer to advanced data science techniques such as machine learning, which is the field where data-driven algorithms are developed.

Data: information (here, that can be used by machines). This information can be numbers, text, images, etc.

Data science: scientific field that extracts, analyzes and draws results from data.

Model training: the act of training a model by feeding it examples so that it self-learns rules and is able to apply them to new cases.

Input data: the information entered at the beginning of the algorithmic process

Output data: the information obtained at the end of the algorithmic process

(Algorithmic) model: there exists a limited number of mathematical algorithms. A model is the product of a mathematical algorithm being encoded, trained and adjusted by a data scientist to become a computer tool.

Reading recommendations:

- Julia Stoyanovich and Falaah Arif Khan. “What is AI?”. We are AI Comics, Vol 1 (2021) https://dataresponsibly.github.io/we-areai/comics/vol1_en.pdf
- Julia Stoyanovich and Falaah Arif Khan. “Learning from Data”. We are AI Comics, Vol 2 (2021) https://dataresponsibly.github.io/we-are-ai/comics/vol2_en.pdf

Bibliography

Ada Lovelace Institute. (2021, 13 May). *Building public confidence in data-driven systems Findings of the Office for Statistics Regulation review into the 2020 exam results algorithm, and why public confidence in data-driven systems matters*. [Webinar]. https://www.adalovelaceinstitute.org/event/building-public-confidence-data-driven-systems/

Adams, R. (2020, 18 March). *All schools to close from Friday; GCSE and A-level exams cancelled – UK Covid-19, as it happened*. The Guardian. https://www.theguardian.com/politics/live/2020/mar/18/uk-coronavirus-live-boris-johnson-pmqs-cbi-urges-government-pay-businesses-directly-saying-350bn-loangrant-package-not-enough?page=with:block-5e7261318f088d7575595edc#block-5e7261318f088d7575595edc

Adams, R.; Elgot, J.; Stewart, H.; Proctor, K. (2020 19 August). *"Ofqual ignored exams warning a month ago amid ministers' pressure"*. The Guardian. https://www.theguardian.com/politics/2020/aug/19/ofqual-was-warned-a-month-ago-that-exams-algorithm-was-volatile

AlgorithmWatch. (2020). *Automating Society Report 2020.* *https://automatingsociety.algorithmwatch.org/*

BBC News. (2020, 14 August, *A-levels: Labour call for government U-turn over 'exams fiasco'.* https://www.bbc.com/news/education-53776938

Bennett, R. and Steven Swinford. (2020, 17 August). *Gavin Williamson apologises as he backs down on A-level and GCSE results*. The Times. https://www.thetimes.co.uk/article/gavin-williamson-apologises-as-he-backs-down-on-a-level-and-gcse-results-05zp3jvl3

Bennett, S. H. (2020, 20 August). *On A Levels, Ofqual and Algorithms.* Sophie Bennett's blog. https://www.sophieheloisebennett.com/posts/a-levels-2020/

Direction under S 129(6) of the Apprenticeships, Skills, Children and Learning Act 2009. (2020 21 March). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/877611/Letter_from_Secretary_of_State_for_Education_to_Sally_Collier.pdf

Duncan, P. (2020, 13 August). *A-level results day 2020 live: 39.1% of pupils' grades in England downgraded - as it happened*. The Guardian. https://www.theguardian.com/education/live/2020/aug/13/a-level-results-day-2020-live-students-teachers-government-ucas-mock-exams-triple-lock-nick-gibb?page=with:block-5f351f6b8f08899e2e66d6b5#block-5f351f6b8f08899e2e66d6b5

Elish, M.C. and Elizabeth Anne Watkins, Data & Society. (2020). *Repairing Innovation: A Study of Integrating AI in Clinical Care.* https://datasociety.net/wp-content/uploads/2020/09/Repairing-Innovation-DataSociety-20200930-1.pdf. p.11

Esson, G. (2020, 4 August). *Scotland's results 2020: How grades were worked out for Scottish pupils.* BBC News. https://www.bbc.com/news/uk-scotland-53580888

Foxglove. (2020, August 12). *Press release: UK: Legal action threatened over algorithm used to grade teenagers' exams.* Statewatch.org. https://www.statewatch.org/news/2020/august/uk-legal-action-threatened-over-algorithm-used-to-grade-teenagers-exams/

Freedman, S. [@Samfr]. (2020, August 13). "Nevertheless the algorithm was (as I said yesterday) inevitably going to hit outlier students who were at the top of…" [Tweet]. Twitter. https://twitter.com/Samfr/status/1293979304179769346

Hao, K. (2020, 20 August). *The UK exam debacle reminds us that algorithms can't fix broken systems.* MIT Technology Review. https://www.technologyreview.com/2020/08/20/1007502/uk-exam-algorithm-cant-fix-broken-system/

Green, B. Z. (2019). *The Smart Enough City: Putting Technology in Its Place to Reclaim Our Urban Future*. MIT Press.

Harkness, T. (2020, 18 August). *How Ofqual failed the algorithm test*. Unherd. https://unherd.com/2020/08/how-ofqual-failed-the-algorithm-test/

ICO. (2020, 14 August). Statement in response to exam results. https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2020/08/statement-in-response-to-exam-results/

Jist Studios, BBC Ideas. (2019, 19 September). "What exactly is an algorithm?" [Video]. https://www.bbc.co.uk/ideas/videos/what-exactly-is-an-algorithm/p07nw8ny

Jones, E. and Safak, C. (2020, 18 August). *Can algorithms ever make the grade?.* Ada Lovelace Institute's blog. https://www.adalovelaceinstitute.org/blog/can-algorithms-ever-make-the-grade/

Kolkman, D. (2020, August 16). *"F**ck the algorithm? What the world can learn from the UK A-level grading algorithm fiasco".* LSE Impact Blog. https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/

Lee, G. (2020, 14 August). *Did England exam system favour private schools?*. 4 News FactCheck. https://www.channel4.com/news/factcheck/factcheck-did-england-exam-system-favour-private-schools

Leufer, D. (2020). *Myth: AI can solve any problem.* AI Myths. https://www.aimyths.org/ai-can-solve-any-problem

Leufer, D. (2020). *Myth: the term AI has a clear meaning*. AI Myths. https://www.aimyths.org/the-term-ai-has-a-clear-meaning

Merrick, R. (2020, 13 August). *Boris Johnson insists A-level results are 'robust' as he shrugs off protests from teachers and pupils.* The Independent. https://www.independent.co.uk/news/uk/politics/boris-johnson-level-results-protests-teachers-pupils-gavin-williamson-a9669131.html

Murkett, C. (2021, 25 February). *Prepare for the next A-level fiasco*. The Spectator. https://www.spectator.co.uk/article/prepare-for-the-next-a-level-fiasco

Ofqual. (2020). *Consultation: Exceptional arrangements for exam grading and assessment in 2020.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/879627/Exceptional_arrangements_for_exam_grading_and_assessment_in_2020.pdf

Ofqual (2020, 13 August). *Executive summary: Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909035/6656-2_-_Executive_summary.pdf. pp. 1-2

Ofqual (2020, April 3rd). *How GCSEs, AS & A levels will be awarded in summer 2020.* Gov.uk. https://www.gov.uk/government/news/how-gcses-as-a-levels-will-be-awarded-in-summer-2020

Ofqual. (2020). *Research and Analysis: Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: interim report.* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909368/6656-1_Awarding_GCSE__AS__A_level__advanced_extension_awards_and_extended_project_qualifications_in_summer_2020_-_interim_report.pdf

Office for Statistics Regulation Authority. *Ensuring statistical models command public confidence: Learning lessons from the approach to developing models for awarding grades in the UK in 2020. Executive summary.* https://osr.statisticsauthority.gov.uk/publication/ensuring-statistical-models-command-public-confidence/

Ofqual's Strategy Risk and Research Directorate (2020). *Research and Analysis: Equality impact assessment: literature review*.

O'Neil, C. (2016), "Bomb parts: What is a model?" and "Shell shocked: My journey of disillusionment". In *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.*

Haines, T. *A-Levels: The Model is not the Student*. Tom SF Haines' website. http://thaines.com/post/alevels2020

Simonite, T. (2020, 10 July). *Meet the Secret Algorithm That's Keeping Students Out of College.* Wired. https://www.wired.com/story/algorithm-set-students-grades-altered-futures/

Smoke, B. [@bencsmoke] for Huck [@huckmagazine]. (2020, August 16). [Tweet, video]. Twitter. https://twitter.com/HUCKmagazine/status/1294985562106015750

Suresh, H. and John Guttag. (2020). *"A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle", Proceedings from the 2020 conference on Fairness, Accountability and Transparency of Machine Learning*, Barcelona.

Stoyanovich, J. and Falaah Arif Khan. (2021). What is AI?. We are AI. https://dataresponsibly.github.io/we-are-ai/comics/vol1_en.pdf

Szymielewicz, K.; Foryciarz A.; Leufer, D. (2020 January 17). *Black-Boxed Politics: Opacity is a Choice in AI Systems*. Medium [Blog]. https://medium.com/@szymielewicz/black-boxed-politics-cebc0d5a54ad

UCAS. (2021). Filling in your UCAS undergraduate application. UCAS. https://www.ucas.com/undergraduate/applying-university/filling-your-ucas-undergraduate-application

Weale, S. and Stewart, H. (2020, 17 August). *A-level and GCSE results in England to be based on teacher assessments in U-turn.* The Guardian. https://www.theguardian.com/education/2020/aug/17/a-levels-gcse-results-england-based-teacher-assessments-government-u-turn

Wylie, B. (2018, 13 August). *Searching for the Smart City's Democratic Future*. Centre for International Governance Innovation. https://www.cigionline.org/articles/searching-smart-citys-democratic-future/