

RESEARCH ARTICLE

Fairness in tension: A socio-technical analysis of an algorithm used to grade students

Merlin Tieleman 

Political science, University of Liège, Liège, Belgium
Email: merlin.tieleman@uliege.be

(Received 30 October 2024; revised 7 March 2025; accepted 16 March 2025)

Abstract

In response to the emergence of COVID in England in 2020, the government declared a suspension of face-to-face education. To deal with the cancellation of exams, then minister of education Gavin Williamson assigned the Office of Qualifications (OfQual) with devising a method for awarding grades to students based on the standards they would have met had the examinations not been cancelled. As a response, OfQual implemented a prediction matrix based on the resources available to them: this system is widely referred to as the OfQual algorithm. However, as the points were delivered on 13 August 2020, disputes arose leading to the cancellation of the algorithm. This paper will focus on the OfQual algorithm as a particularly relevant case to highlight the tensions around the notion of fairness in the implementation of such systems. Drawing on Science and Technology Studies, I will start by opening the black box of the OfQual algorithm. This will, in turn, allow me to identify the conflicting accounts of what is considered as fair in such system and how such accounts were inscribed within this algorithm, questioning what it means for such system to be fair in contexts marked by inequalities.

Keywords: actor network theory; algorithmic discrimination; fairness; social inequalities in education; sociology of assessment

1. Introduction

In response to the emergence of COVID in England in 2020, the government declared a suspension of face-to-face education. The announcement was made by the then Secretary of State for Education, Gavin Williamson (2020), and it was decided that the planned exams would not take place. To manage the cancellation of the exams, Williamson commissioned the Office of Qualifications (OfQual) to develop a method of awarding grades to students based on the standards they would have achieved had the exams not been cancelled. OfQual was therefore asked to develop a system for awarding grades for two of the most important examinations in the English education system: CGSE and A-Level. This paper will focus on the latter, as it serves as the basis for students to gain admission to their chosen higher education programmes.

OfQual developed their system relying on two main resources (OfQual, 2020a). First a ranking of students provided by each school, known as ‘Centre’s Assessed Grades’ (CAGs). Then the historical performance of each student and school. Following the government’s commendation, OfQual thus created a prediction matrix in just four months. This matrix was designed to reduce the feared

grade inflation caused by the cancellation of exams: this system is widely referred to as the OfQual algorithm.

However, when grades were handed in on 13 August 2020, controversy erupted. The standardisation process introduced by OfQual had resulted in around forty per cent of students' grades being lower than their GACs. This led to protests on the streets of London and, on 17 August, the scrapping of the standardised grades, which were replaced by the higher grades between the GACs and the calculated grades. The abolition of grades in turn led to the grade inflation that the Department for Education wanted to avoid in the first place. It also meant that university enrolment processes had to be reopened, with the logistics of such a manoeuvre left to the universities. However, by the time the process was completed, most study programmes had already been booked. As a result, the government had to lift the cap on student numbers at universities to allow new students to enter. On 15 September, Gavin Williamson was dismissed as Education Secretary as Prime Minister Boris Johnson reshuffled his team after the first phase of the pandemic Reshuffle, 2021).

In this article, I will look at the creation, implementation and undoing of the OfQual algorithm as a particularly relevant case for exploring notions of fairness in algorithms. To do this, I will begin by opening the black box of the OfQual algorithm from a sociotechnical perspective, highlighting the ways in which it has been imbued with the values of its societal context. In doing so, I will highlight how the OfQual algorithm has emphasised the maintenance of standards rather than the ability of students to self-determine, leading to a situation where fairness lies in repeating existing inequalities rather than combating them. To do this, I will explore the work of science and technology researchers Andrew Selbst, Boyd, Friedler, Venkatasubramanian and Vertesi (2019) who have identified five key moments in the development of automated systems where developers are at risk of falling into pitfalls. Throughout the identification of these five pitfalls, I will highlight how they highlight different accounts of what is considered fair in the case of OfQual. These conceptions of fairness will then be put into perspective with the technical possibilities of the system, its *script* (Akrich, 1987). Identifying both the accounts of fairness and their relation to the *script* allowing me to underline how different negotiations happening in the making of the algorithm shaped its constitution.

2. Opening the black box of algorithms

As Europe's Artificial Intelligence Act encourages states to put in place *regulatory sandboxes* to provide with 'lighter' legal framework to allow innovation to prosper (Laurent, 2024), it becomes crucial to be able to study the way values are embedded within such systems. The OfQual algorithm is an especially relevant case to do so as it was implemented in a context of emergency induced by the COVID pandemic where the ministry of education had to act in a timely and cost-effective manner. This led to the confection of a system in fringe of the existing legislation in what could be qualified as an *innovation monster*: 'demonstrable objects of public admiration or public fear, which do not exist without being displayed [and that are] connected with anticipations of progress or catastrophe, and cause pervasive uncertainties about their future evolutions' (Laurent, 2024). Studying the way such system is built from a socio-technical perspective, considering the embedment of social values in the making of algorithms, thus become crucial to understand the potentially harmful effects of those *innovation monsters*. Yet studying algorithms from a socio-technical perspective is no simple task. Approaches that are being developed by scientists in critical algorithms studies to open such black boxes are relatively new and do not always come with fully developed methodologies.

2.1. Approaching algorithms as sociotechnical artifacts

To open algorithms with a socio-technical approach, we can turn to Science and Technology Studies (STS). Indeed, opening black boxes is a prominent approach in STS. In a seminal article, Langdon Winner (1993) writes that 'the term black box in both technical and social science parlance is a device

or system that, for convenience, is described solely in terms of its inputs and outputs.’ This echoes some of his earlier work in which he underlines how technologies can be political (Winner, 1986), both through the values instilled in them during their making and through the social practices required to make them work. Black boxes can thus be considered as placeholders for complex and opaque systems, but this does not mean that they are unopenable.

One way to open them is to focus on the way they are challenged through controversies. This is the hallmark of the *actor network theory* (ANT) developed by STS scholars Madelaine Akrich, Michel Callon and Bruno Latour (Akrich, Callon & Latour, 2006). With their approach, they suggest focusing on the way network of actants – both humans and non-humans – are built, maintained, and changed through various negotiations. Doing so, they encourage researchers to identify *obligatory passage* points through which the objectives of all the *actants*¹ meet. To open black boxes, they argue that we should study the way actants are changed, or *translated*, through various negotiations to become an *assemblage*, a temporary stabilisation of a network where every actant finds its interest. Studying controversy around black boxes is thus a way to identify what is necessary for such network to stay in place.

Then, Akrich (1987) argues that as part of this negotiation process, innovators *inscribe* their own vision into their end product. The inscription leading to the implementation of a *script* within a socio-technical system, a framework of action defined by its makers but also by the space in which it is put in place. In other words, the *script* of an object consists of what it prompts users to do, regardless of how it was intended to be used by its makers. The work of *de-scription*, focusing on the mechanisms of adjustments between users as imagined by the makers and real-world users, then become crucial to identify the script of a socio-technical system. Finally, Akrich (1987) adds that as technological objects stabilise, they are turned into instruments of knowledge. If a sociotechnical system successfully becomes a black box, it is then able to transform sociotechnical facts – resulting from various negotiations – into facts that are perceived as true *de facto*.

2.2. Approaching the OfQual algorithm in practice

Building on this literature from STS, some authors started to open the black boxes of algorithms with a socio-technical perspective, paving the way for critical algorithms studies. Authors as Angèle Christin (2020) for example identifies how the opaqueness of algorithms is socially enacted and encourages researchers to study how algorithms produce a *refraction* of our societies by assembling a skewed reflection of it. Nick Seaver (2017) proposes to see algorithms as being entanglements of social practices that are culturally enacted by people who engage with them. Consequently, algorithms are part of a cultural stream and not solely an artefact interacting with it anymore. In practice, Seaver’s (2017) approach suggests that algorithms should thus not be seen as apart from society as it is not possible to isolate them from their societal context. Others such as Rob Kitchin (2017) propose ways to engage with them in practice by dismantling piece by piece the socio-technical assemblage surrounding algorithms.

As a data collection strategy, Seaver (2017) suggests practising what he calls *scavenging* by collecting secondary and sometimes tertiary sources to grasp the algorithms that are often hidden behind complex codes or secrecy from their developers. Doing so, one should not reject resources based on their intelligibility. Indeed, the inaccessibility (both in term of difficulty to comprehend, but also to access in the first place) of sources is also a part of the analysis as it is a major aspect of black boxes (Christin, 2020). Collecting codes, gargantuan datasets, and specialists’ reports is thus also part of such research. These resources in turn allow researchers to deconstruct the code step by step, by including all comments and documentation available (Kitchin, 2017). It is these scavenged resources that in turn grant access to the identification of the step-by-step process that transform an input into

¹The notion of *actant* is preferred here to its corollary concepts of *actors* as the term *actant* take into account the diverse form of actors, humans and non-humans (the code, the computer, etc.) at play in our case.

an output within an algorithm. The study of said process allowing us to engage critically with the algorithm to identify how their *script* is imbued with certain values.

In the case of OfQual, as it was part of a public policy in order to overcome the cancellation of exams, there was a commitment to be as transparent as possible in order to build confidence in the system from the public (OfQual, 2020a). As a result, a lot of documents were produced explaining the choice made by OfQual and the inner workings of their algorithm. Most of the analysis will thus be built on the literature made available by OfQual, as well as a parliamentary inquiry on the impact of COVID-19 on education and children's services (2020). Both those resources regrouping perspectives from various stakeholders (developers, executive directors, students, teachers, etc.). With those, I will retrospectively rebuild, step by step, the different negotiations that happened while making and implementing the algorithm. Following on classical STS literature (Akrich, 1987; Bijker, 1995; Callon, 1984), I will pay particular attention to how the different negotiations happened within the making and implementation of the system that led to its contested outcome. Following Bloor's (1976) *principle of symmetry* – that encourages us to avoid studying a past controversy following a discourse *a posteriori* – I will avoid identifying the OfQual algorithm as being neither right nor wrong. I will thus avoid perceiving it as being inherently unfair against some, but I will instead identify how different discourses around fairness emerge in its making.

2.3. Fairness in algorithms

The STS perspective offers great possibilities for approaching algorithms as socio-technical artefacts, imbued with values derived from their production. But in order to understand how the notion of fairness is approached in the creation of the algorithm in question here, we need to take a brief detour into theories of fairness in algorithms.

As STS scholar Bilel Benbouzid (2022) points out in his analysis of Fair Machine Learning (or FairML) as a field of study, the ways in which fairness is considered have evolved over time, with sometimes conflicting views. He highlights a tension within FairML between the methodological realism of engineers and the constructivism of social scientists. In the former case, algorithms are seen as capturing an objective reality or ground truth, echoing the work of STS scholar Folrian Jatón and Bowker (2020). In the constructivist approach, engineers themselves define the criteria for what is 'fair,' using metrics and formal rules. As Benbouzid (2022) argues, the field of FairML offers a rare case where the two approaches coexist in a form of *applied realism*.

Having identified this tension, he then explores different, sometimes conflicting, measures to account for different approaches to fairness. First, statistical fairness, to ensure that different groups are treated equally. Then procedural fairness, which ensures that decision-making processes follow certain transparent and fair rules. Finally, individual fairness, which seeks to treat similar individuals in a similar way.

At another level, authors such as Melanie Smallman (2022) show that there are conflicting accounts of fairness due to the multiple levels at which fairness is defined. Looking at existing frameworks for regulating ethics in AI in welfare, she argues that 'a rights-based approach has focused our gaze on the individual' (Smallman (2022, p. 13) overcome this shortcoming, she suggests looking at fairness from different perspectives). She suggests six scales: individual, group and community, institutional, national, global and over time. Approaching fairness from this multi-level perspective allows us to better anticipate the far-reaching consequences of the implementation of algorithms in our societies.

Authors such as Benbouzid (2022) and Smallman (2022) thus offer great accounts of how fairness can be a contested matter within critical algorithms studies. However, in order to explain these conflicting accounts of fairness in the development of algorithms, I will build on the work of STS scholars Selbst et al. (2019) to highlight how the OfQual case demonstrates different accounts of fairness in the development of algorithms. In their work, they identify five key moments in algorithm development that developers can fall into. Traps that lead to abstraction errors, where the reality represented

by an algorithm is no longer true to its ground truth. These five moments are presented as *traps* that developers can fall into.

First, the *solutionism* trap, which invites developers to consider when to design and whether an algorithm is needed at all. The second is the *ripple effect* trap, which asks us to consider the specifics of each scenario when implementing automated systems. The third is the *formalist* trap, which calls to avoid reinforcing existing inequalities or introducing new ones. The next is the *portability* trap, which reminds us to include the perspectives of all relevant social groups in each situation. Finally, the *framing* trap invites developers to consider the whole system into which the algorithm will fit.

In this paper, I will begin by opening the black box of the OfQual algorithm by identifying how various negotiations happening in its making and implementation shaped its *script*. Following on Seaver's (Seaver, 2017) technique of *scavenging*, I will build on reports made by OfQual and the government, complementing those with interviews with OfQual developers. Then, I will identify how the different accounts of fairness are intertwined throughout the five moments identified by Selbst et al. (2019) to underline the *script* (Akrich, 1987) of the algorithm.

3. The OfQual algorithm

The process of opening black boxes of algorithms is a complex one. In this section, I will provide a comprehensive analysis of the development of the OfQual algorithm from a socio-technical perspective to highlight the numerous ways in which this technical system was intertwined with society. This will then allow me to identify the diverse values that were embedded within the algorithm. But first, I will provide with a brief explanation of the context in which the algorithm was built.

3.1. Education in England

To better understand the context in which the algorithm was implemented, I first propose to trace the path of a grade assignment through the English educational system in a year without the algorithm. This will then allow us to understand the steppingstones on which the OfQual algorithm was built, and how the algorithm introduced in 2020 has reshaped the assessment pathway.

The first set of individuals we encounter are the students being assessed. In this case, these are all the students who took their A-Levels in England in the year 2020. A-Levels are exams taken at the end of secondary education by students in England. Typically, students choose between two or three A-Level subjects (i.e. mathematics, geography, and economics) in a centre of their choosing, depending on the requirements of the colleges they wish to attend. This means that around 700,000 A-Levels needed to be taken in 2020 for about 250,000 students by summer 2020 (OfQual, 2020a). The A-Levels, once passed and graded, are used by students to access their desired programs within higher education, the highest the better. The grade ranged from A+ to E, with the grade U (for ungraded) given to students depending on the percentage of students reaching each mark in their exams (the top 10% receive A+, top 20% gets A, then B and so on and so forth).

The next actors in the grading process are the teacher. Their aim is to prepare students for their final exams through teaching and assessment throughout the year. These teachers are part of schools and colleges, which OfQual calls 'centres' (OfQual, 2020a) as the term better encompasses the wide variety of educational formats in England. A student wishing to take an A-level will therefore join a centre that offers A-levels in the subject they wish to take. Each centre is run by a headmaster, who played an important role in approving the CAGs for the 2020 evaluation (OfQual, 2020a).

Next in the graduation process are the exams boards, which each centre asks to provide certified A-Level exams. There are currently five exam boards in England: AQA, OCR, Pearson, Cambridge and WJEC. In many cases, the different boards offer A-Levels in the same subject, with centres having to choose between different boards to provide them with A-Level assessments. According to OfQual

(2020c), centres may choose one board over another based on the form of the assessment, the support service or the specifications of each centre.

Each exam board provides the requested examinations and ensures the correction of the said examinations. They then provide students with their grades via a standardised national platform. For education experts Denis Opposs & al. (2020), the exam boards in the English educational system are typical of what they call a *quasi-market system*. According to them, the delegation of assessment to private third-party organisations is based on the belief that competition between different boards will push them to raise the overall level of education, as each board wants to have the best performing students. The system has been contested in England by politicians from various backgrounds, but it is still in use to this day (Opposs et al., 2020).

The list of subjects that exam boards are allowed to award is set by OfQual based on contents set by the Department for Education. OfQual is specific to England, and it is the only regulatory office for education under the direct authority of the UK government. Indeed, parliaments of Scotland, Wales and Ireland are responsible for their matters in education. Exam boards are a consequence of the peculiarities of the quasi-market approach promoted in England, as the other parliaments provide the qualification themselves without referring to third-party administrations (Opposs et al., 2020). OfQual objective is to regulate the different exam boards to ensure comparability over time. It is a non-ministerial governmental organisation, which means that they must follow the recommendations of the Department for Education, but they can do so with a certain degree of freedom.

3.2. Settling for the algorithm

The English education system is one in which examinations are outsourced to independent organisations following a logic of competition, in our case examination boards. Such system lies on the belief that as each board wants to have the highest achieving students; they uplift the general level of education through competition between them (Opposs et al., 2020). To maintain such system in place, grades must be comparable not only between different exam boards but also through time. As the OfQual algorithm was introduced in such a context, standardisation was an important aspect of its development.

Such constraint is acknowledged by the Department for Education who asked OfQual ‘that qualification standards are maintained, and the distribution of grades follows a similar profile to that in previous years’ (Williamson, 2020). This meant that OfQual had to come up with a way to standardise results for A-Levels to avoid them being too high or too low compared to the year-on-year average. To do so, OfQual put in place a process to smooth the CAGs so that the general grade distribution would be in line with historical distribution. To achieve their goal, they agreed on the following set of objectives that they presented in their interim report.

“The confirmed aims of the standardisation process are therefore:

- i. *to provide students with the grades that they would most likely have achieved had they been able to complete their assessments in summer 2020*
- ii. *to apply a common standardisation approach, within and across subjects, for as many students as possible*
- iii. *to protect, so far as is possible, all students from being systematically advantaged or disadvantaged, notwithstanding their socio-economic background or whether they have a protected characteristic*
- iv. *to be deliverable by exam boards in a consistent and timely way that they can quality assure and can be overseen effectively by Ofqual*

- v. to use a method that is transparent and easy to explain, wherever possible, to encourage engagement and build confidence” (OfQual, 2020a).

To achieve these goals, OfQual had different resources at hand. First, as predictions about the general average of students are made each year to predict the grade distribution² (Kelly, 2021; Ozga, Baird, Saville, Arnott & Hell, 2023), OfQual already had access to a lot of historical data about the history of each centre. Indeed, they knew the final grades for each student’s A-Level in the years before the algorithm as well as most of the results of the CGSEs of students in the 2020 cohort.

On top of that, OfQual announced in a report from the 3rd of April 2020 (2020a) that they were going to ask centres to predict students’ grades and to rank them. To do so, centres were instructed to give a grade for each student for each A-Level and to rank students within each subject with the highest attaining student on top. Those rankings are the CAGs which I introduced earlier. They asked for these resources before knowing what they were going to do with them as they wanted to act in a timely manner.

Yet, they relied on the belief that the predictions of teachers would not be sufficient as they would be higher than grades students usually achieve. Indeed, OfQual (2020b) used previous studies on grade prediction which suggested that teachers tended to be less precise when asked to grade students in general (*absolute accuracy*) than when they were asked to sort students in comparison with one another (*relative accuracy*). They also used tests on their historic data to support that *absolute accuracy* was less optimal for their objectives (OfQual, 2020a). According to them, teachers were imprecise in about half of the cases. And most of those times, the teachers were too optimistic. Using those sources as a reference, OfQual argued that using grades predicted by teachers through absolute accuracy directly would have led to an inflation in grades for the year 2020. This was verified as most of the CAGs given by teachers were indeed above the national average. But to say it is only due to the wrong appreciation of teachers might be a hasty conclusion. Indeed, as teachers were aware that the grade they gave were going to be standardised, it might be possible that they raised their students’ grades above their expectations.

OfQual (2020) also argued that using CAGs directly would have led to inequalities between centres and students. As it was not possible for OfQual to properly form teachers to produce accurate standards in such a timely manner, there would be too many disparities between centres and variation between years would be too high and lessen the values of higher grades for the 2020 cohort. This led OfQual to consider a way to standardise the results given by teachers through the CAGs to fit with the historical distribution of grades.

It is important to note here that OfQual put in place an online survey to do public consultation around the objectives their system should pursue (OfQual, 2020c). Those happened between the 15th of April and the 22nd of May and the 8th of June and receive feedback from 1,939 students amongst 12,623 responses, a low margin regarding all A-Levels being awarded (OfQual, 2020d). Moreover, the matter of identifying whether OfQual or the Ministry of Education is responsible for deciding to implement the approach that ended being used is contested (Ozga et al., 2023). Calling this case ‘the OfQual algorithm’ might therefore be a shortcut for more complex issues of relation of power within the English administration, but those are out of the scope of this sole paper.

3.3. The algorithm in action

OfQual had three main resources at hand: centres’ historic grade distribution, students’ previous achievements, and the CAGs. In order to provide students of the 2020 cohort with grades with such data available, OfQual tested eleven prediction models (OfQual, 2020a). In the end, they chose to use

²In a year with examination, predictions are used to decide to what grade corresponds each mark within each exam board and within each subject.

a model called the *Direct Centre Approach* (DCP) which focused mainly on the historical distribution of grade within each centre, favouring the specificities of centres over the national average or the students’ histories. They chose the DCP approach as it was the one affording the most reliable results whilst also being cost effective – both in term of money and time – to put in place (OfQual, 2020a). The algorithm was later made public by OfQual in an extensive report explaining the way their model works.

The algorithm starts by producing a first prediction of the grade distribution for an A-Level in each centre. This prediction is made based on the historical grade distribution of each centre for each A-Level, considering the average grade distribution of a centre over the last three years. The relatively low number of years for the comparison is due to a change in regulation in 2016 meaning that the test could only be made with the data from 2016 to 2019. The algorithm thus had to follow the same time-period to ensure its accuracy compared to the testing. One downside of this distribution was that the fast change in grade distribution (either upward or downward) of some centres was not considered by the algorithm.

The algorithm then predicts a grade distribution for the second time based on students’ previous achievements. Of course, when processed by computers, this step happens at the same time as the first one, but for the clarity of the explanation, it is easier to break it into steps here. During this part, not all students are being taken into consideration. Indeed, some students cannot be matched to their previous achievement by the algorithm, for example students coming from abroad. To ensure that the general distribution of grades is not affected too much by this limitation, the formula developed by OfQual weights the proportion of unmatched students with the national grade distribution. This means that in a classroom where most students were unmatched, the predicted distribution would be closer to the national one.

The model then adjusts the initial distribution of grades to be more in line with previous achievements and the history of each centre and to produce a *prediction matrix* that is in line with both the historical grade distribution of each centre and the previous achievement of students. The students are then matched to the *prediction matrix* based on their ranking within the CAGs. For example, if there are 12% students predicted to get a grade A+, the first 12% of the ranking get the grade A+. To avoid having students between two grades,³ marks are calculated following the predicted distribution and the history of each centre. The student’s grade is then settled based on that predicted mark.

The working of the OfQual algorithm can be simplified with the following representation (Figure 1):

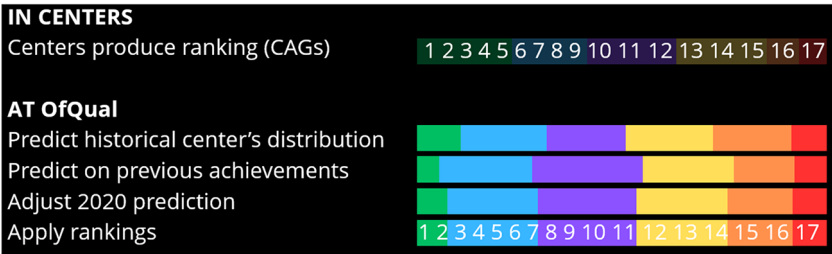


Figure 1. Simplified representation of the OfQual algorithm, realised by the author.

The way the model works means that the students’ calculated grades strongly relied on the proportion of students that achieved each grade from 2017 to 2019 within each centre, giving more weight to the history of each centre than to the previous achievement of students and their CAGs. That is a direct consequence of the government’s injunction that I underlined earlier. Indeed, this was made to

³This could happen as, for example, if 12% of students were getting an A* in a centre of 20, which means that between 2 and 3 students could get that grade.

ensure that the general average of each centre, and thus the national average, would not change too much.

3.3.1. *Exceptions within the algorithm*

While the algorithm was applied to all A-Levels, OfQual (2020) was not able to fit all students within their model. Indeed, as the model relied strongly on the history of centres, it meant that students isolated within the educational system could not be assigned with accurate results. Such situations generally occur with private candidates – doing their education independently from the centres but still taking their A-Levels in centres – and candidates within small cohorts.

Regarding the first category, students refer to a centre only for the examinations but not for their education. To prevent those students from disturbing the rank order of the students who attended the centres, they were removed from the ranking before applying the DCP model. Their grade was then calculated by seeing how they compared to students around them in the initial ranking. In cases where they were between two grades, or if they were first or last in the ranking, their grade was defined by their CAG. Yet, interviews with students in this situation reveal that it sometimes happened that they received no grade as the centre they relied on believed that giving them a rank would still disturb the calculated grade of their students.

Another exception was made for students in small cohorts. As those do not have enough history to build a precise and consistent distribution of grades, it was not possible to apply the DCP model. OfQual thus opted in favour of giving the students of those centres their CAGs directly as using their model would not be representative of the prior attainments from students. OfQual measured that to keep in line with the general distribution of grades over the years, centres under fifteen could be handed their CAGs directly. All centres with entries over fifteen students by A-Level going through the standardisation process.

3.4. *Testing the algorithm for equity*

As OfQual, following on the incentives of the government for education, successfully made an algorithm reproducing historical grade distribution at a national level, it raised the question of the presence of inequalities that could be imbued within this new process. In their interim report (OfQual, 2020a), OfQual explained how they tested their system regarding equity. It is important to note here that OfQual's focus was on the mitigation of biases. Indeed, as they put it, 'the analyses conducted show no evidence that this year's process of awarding grades has induced bias. Changes in outcomes for students with different protected characteristics and from different socio-economic backgrounds are similar to those seen between 2018 and 2019.' (OfQual, 2020a). But how did they come to this conclusion? In addition, is debiasing the solution toward a fair algorithm in this case?

To identify whether their model had a significant impact on social equity, OfQual identified a set of variables regarding students' backgrounds. Two of the variables were taken from exam boards directly: gender and prior attainments. And five variables were taken from the national pupil database (NPD): ethnicity, major language, special educational needs, free school meal, social economic status. Based on these variables OfQual then analysed how the gaps in grade distribution between the different categories were impacted by their model. They did such analysis not only for each category independently, but for students crossed categories (gender crossed with ethnicity for example).

Doing so, OfQual ensured that their model accurately reproduced the historical grade distribution within the English educational system. As the algorithm became a controversial issue after the delivery of grade, partly due to accusation regarding the unequitable nature of the model, OfQual made their whole dataset as well as the algorithm available for independent researchers to study the model. Despite the accusations of the protesters, scholars from various disciplines studying the algorithm all argue that the algorithm in fact did not discriminate against any protected categories identified by OfQual. Indeed, those studies showed that 'it is unlikely that students belonging to any of the

combinations of protected characteristics considered here were worse off in terms of their teacher-assessed grades when compared to what they would have been likely to have received had COVID interruptions not occurred' (Magowan, 2023).

Indeed, looking at the tables given by OfQual (2020), we can see for example that students benefiting from free school meals, an indicator of poorer socio-economic status, were 6.7% less likely to obtain a grade A or above than students who did not benefit from it in 2018. They were 7.1% less likely to achieve such a grade in 2020. Overall, students from a lower socio-economic status were 7.3% less likely to obtain a grade A or above in 2020, a number in line with previous years. Such a repetition of historical differences in the distribution of grades between students holds true for each background variable. Overall, the algorithm successfully reproduced historical grade differences between the tested protected categories that were identified earlier.

Of course, the composition of each of these categories could be questioned, as the sorting of students in said manner had a direct impact on what was tested regarding inequalities. For example, the division of the gender category between male and female students, removing the *other* category due to the 'very low numbers' (OfQual, 2020a) meant that OfQual had no information on the impact of their model on transgendered students for example. Likewise, the variables regarding ethnicity could also be questioned with OfQual doing their analysis on six ethnic groups (Asians, Black, White, Mixed, Other, and Unclassified), and on one *minor* ethnic group as qualified by the NPD: the Chinese, usually a sub-group of the Asian ethnic group.⁴

Then, OfQual underlined in their report that there is no indication of discriminatory biases against protected categories in general. Doing so, they refer to the 2010 equality act regarding protected categories that should not be discriminated against. Those categories are as follows: age, gender reassignment, civil status, pregnancy, disability, race (colour, nationality, ethnic and origins), belief, sex, and sexual orientation. Whilst some categories might not be relevant for the case at hand, like age for example, not all protected categories have been tested by OfQual. With a model relying heavily on teachers' ranking of students, we might ask whether students from the queer community might have been discriminated against by their teachers, coming worse off than they could have with the algorithm.

Through their testing process, what OfQual demonstrated is that the points obtained by each protected category identified while testing for inequalities were in line with previous years variations. But what that meant is that attainment gaps between said backgrounds were not only maintained but also institutionalised in an automated system reproducing existing achievement gaps between categories. Indeed, with this system in place, it meant for example that no more than around 17.7%⁵ of students belonging to the black ethnicity group could achieve a grade A or above.

3.5. Contesting the algorithm

Due to the 2020 standardisation model, 41.3% of the CAGs had been tampered with to fit with the historical distribution of their centres. A 39.1% of the total being downgraded one or two grades. As I demonstrated earlier, the results of the calculated grades showed no anomalies in terms of equality between students in comparison to previous years regarding the controlled categories. Its main consequence was to lower the grades that were given by the teachers. The algorithm had thus fulfilled its goal given by OfQual and the ministry of education to provide students with grades in line with the standards.

⁴While this is out of the scope of this paper, questioning the categories set by OfQual could indicate that while it is true that the algorithm was unbiased regarding the tested categories, it might not be true for other sets of individuals.

⁵This number corresponds to the percentage of black students achieving grade A or above in 2020 (17.3% in 2018 and 16.4% in 2019)(OfQual, 2020a).

But on the 13th of august 2020, as students received their grades and OfQual released their interim report explaining the algorithm, contestation appeared. Quickly, *#fuckthealgorithm* was used as a rallying cry for students both online and offline (Benjamin, 2022). On the 16th of August, hundreds of students chanted their rallying cry in front of the Department for Education building with placards stating things such as ‘I’m a student not a statistic’ or ‘poor \neq stupid’ (Benjamin, 2022). The general contestation following the implementation of the algorithm has been studied by many scholars following different disciplines, with two main arguments against the algorithm underlined: first against the results of the algorithm, second about the removal of students in the process.

Authors such as Kelly (2021) argued that protests were stemming from the belief that students from lower-socio economic backgrounds were at higher risk of seeing their grades downgraded. The public enquiry on the impact of COVID-19 on education and children’s services (2020) give us great account of such fear, regrouping a total sum of 454 written evidence around the impact of Covid on education. The following testimony from the enquiry being an illustration of such concern:

*“Despite the fact that we were all accepted to top universities in the UK, including LSE, Oxford, and Cambridge, our dreams are about to get abolished. But the worst part is that it will not be because we did not work hard enough. Instead, it is due to the fact that we all go to a school with low historical grades in all subjects. Even if our teachers gave us 3A*s, we will inevitably be downgraded if Ofqual does not amend its standardisation model.”* (Smith, Hong & Hart, 2020)

Indeed, while it untrue to say that the algorithm induces biases against the tested protected categories in comparison to other years (Magowan, 2023), it is still true that students from low achieving backgrounds and protected groups were bound to achieve lower grades as they were limited to the historical best their background allowed.

This led to new conflicting discourses when the algorithm was revealed, leaving the government with two options: enforcing the calculated grades at all costs without taking into consideration the revendications of students and teachers, or accept the expertise of teachers and apply their CAGs directly. But as they chose the second option, that meant that the algorithm was rendered useless.

However, a study of the marks awarded to pupils in 2020 after the cancellation enables us to test some of the hypotheses put forward by OfQual and the Ministry to implement the algorithm. Indeed, the grades showed that the disparities between centres in fact did not widen but stayed stable regarding the history of each centre. Moreover, as I exposed earlier, there is no evidence that the CAGs induced biases against protected categories identified by OfQual in comparison with previous years (Magowan, 2023; OfQual, 2020a). The main difference brought by the CAGs being the overall higher distribution of A* and A grades. The fact that the general grade distribution between protected categories does not change after the U-turn teaches us that while teachers might tend to over-estimate their students in general, they are still very good at reproducing existing distinction between students.

4. Discussion – abstraction traps and the multiple accounts of fairness

Multiple negotiations took place throughout the making of the OfQual algorithm, eventually leading to its cancellation. Authors studying contestation around the algorithm, such as Benjamin (2022) and Heaton, Nichele, Clos and Fischer (2023) argued that parts of this contestation was due to conflicting accounts of what fairness meant between the ministry of education and students. Throughout the previous section, I demonstrated how negotiations happening in the making of the algorithm shaped its *script* (Akrich, 1987). In this last section, I will thus identify how these negotiations represent different, and sometimes conflicting, accounts of what fairness is.

To do so, I will build on the work of STS scholars Selbst et al. (2019) – who identified five key moments in the development of algorithms – to underline how the OfQual case showcases different account of fairness in the making of algorithms. These moments are *traps* in which developers might fall (Selbst et al., 2019): with fairness considered as separate from its social context. Traps leading to

abstraction error, where the reality being represented by an algorithm is no longer true to what it originally was. The five traps are as follow: the solutionism trap, the ripple effect trap, the formalist trap, the portability trap, and the framing trap. In parallel, I will identify how the different accounts of fairness are intertwined with the *script* (Akrich, 1987) of the algorithm. The identification of the *script*, as well as what it prompts or refrain people to do, will help revealing how its makers envisioned the educational system in the first place.

The first trap identified by Selbst et al. (2019) is *solutionism*, calling the attention of developers towards the necessity to consider when to design. In the case of OfQual, Kelly (2021) argues that, instead of implementing an algorithm, universities could have directly used the actual grades of students for their admission – through continuous evaluation up to the point of courses cancellation. This was the method used in countries such as France and Belgium.

While Gavin Williamson never explained why he chose to direct OfQual towards calculated grades, a plausible explanation for the Ministry of Education's urge to put an algorithm in place rather than relying on continuous evaluation may come from the fact that certificates need to be delivered by exam boards. Indeed, the English educational being a *quasi-market system* (Opposs et al., 2020), it means that have assessment have to be made by private third-party organisations that compete with one another; the exam boards. OfQual was therefore not able to guarantee such system without the boards, nor equality in standards across boards and centres as continuous evaluation practices may vary.

This shift from an objective of evaluation to one of comparability represents a new negotiation around the aims of evaluation in England. Going back to the *ANT*, OfQual being tasked with the role of assessment implies a change in *obligatory passage point*, meaning that the whole system is being negotiated once again. The Ministry's original objective of ensuring comparability between years was changed into the development of a grading system to ensure comparability itself, consequently inscribing the algorithm with a specific script.

This provides us with a first account of what fairness is in the case of the OfQual algorithm where fairness lies on year-on-year comparison between students. This view lies on the need for the Ministry of Education to compare cohort through time as it is a major aspect of the *quasi-market system*. It is with that assumption on fairness, relying on market competition between third-party organisations, that OfQual started to make their algorithm. Doing so, they inscribed their system with the objective of ensuring comparison over time, nudging its development in a specific direction.

The second trap identified by Selbst et al. (2019) is the *ripple effect*, bringing attention to what computer scientist Rob Kling calls *reinforcement politics* and *reactivity* (Kling, 1991). In other words, he highlights the importance of paying attention to the ways new technologies might have a negative outcome on minoritized population or minoritize some as a result. As I showed earlier, OfQual was in fact careful not to produce discriminatory outcomes by minimising the biases produced by their algorithm. But whilst OfQual was right in stating that there is no evidence that their system *introduced* biases (OfQual, 2020a), failing to take into consideration the *ripple effect* meant that they were in fact *institutionalising* achievement gaps.

Indeed, the developers focused solely on the exact reproduction of grade distribution between measured categories. Doing so, they did not consider that it would be impossible for students in said categories to achieve better than their predecessors. By arguing that fairness lies in the impeachment of additionally induced inequalities, it meant that the system put in place in 2020 did not only repeat existing distinction between categories but also made them immutable.

Moreover, the making of the tested categories could also be questioned. Indeed, the process of debiasing made by OfQual only took into consideration some categories of students. Again, a deeper study of the rationale behind the choice of tested categories might help us understand better the way the algorithm had an impact on protected categories. Yet, OfQual needed to be able to link each variable to a student to do their analysis. Whilst this was possible with already linked variables such as sex or ethnicity, it might have been more complicated with variables such as sexual orientation,

which are typically not registered. This of course reveals how the script of the algorithm, inscribed by its makers, lead to a system unable to consider some categories of people.

Taking fairness into consideration in this case reveals the difficulty of automating evaluation while removing the ability for students to prove their worth in an evaluation. This impossibility to account for some students comes in opposition with the argument that fairness lies in the repetition of historical grade to ensure comparability. But as the algorithm was made with the objective of comparability in mind, following the script of the algorithm makes it clear that its developers inscribed it with such vision. This meant that as the algorithm was inscribed with the idea that fairness lies in comparative justice, repeating existing inequalities between groups.

The third trap identified by Selbst et al. (2019) is the *formalism*, with which they invite developers to include the point of view of the different *relevant social groups* in the making of an algorithm. Doing so, it is important to pay attention to the different points of views of the publics at hand. In the OfQual case, I already identified the shift in obligatory passage point from examinations to the model developed by OfQual. Following the incentives from the ministry of education, OfQual removed students from the negotiations taking place during an examination. Doing so, they scrapped those of their status as *relevant social groups* in the negotiations around evaluations. This meant that they were unable to take into consideration what matters to those groups, leading to contestation in the end.

Moreover, by putting in place the algorithm as an obligatory passage point instead of examinations, OfQual and the government had rendered normal the existing attainment gaps within the educational system. It was thus not possible for students to fight against those by performing in their examinations, given that the legitimacy of the system relied on its self-reproduction without the implication of students. The shift of obligatory passage point had opened new negotiations around what should be the content of the evaluation, not including the students in the discussion. By failing to include students in the grading process, OfQual and the ministry revealed that they did not consider them as relevant for the evaluation process, once again putting fairness in the hands of year-on-year repetition of inequalities.

The fourth trap identified by Selbst et al. (2019) is *portability*, which calls for attention to the specifics of each user's script. Doing so, developers should be careful about the peculiarity of the context in which their algorithm is put in place. In the case of OfQual, the algorithm was made to be applied on a national level, overlooking the peculiarities not only of each centre, but also of each student's individual trajectory. This had a direct consequence on the difference of treatment between students, as some centres were assigned the CAGs directly instead of the standardised grades. At the same time, other students received abnormally low grades in comparison to what they would have achieved as the algorithm flattened their progression.

According to Kelly (2021), the choice to apply the CAGs directly to small cohorts gave an unfair advantage to wealthier independent centres.⁶ Indeed, he argued that small centres receiving their CAGs directly were mainly richer, fee-paying centres, that had already selected students based on their prior attainments. However, one should note that some pupils from such fee-paying high-profile centres also saw their grade being reduced due to the process of standardisation because of a lack in historical data (Hussain, 2020).

Yet, the choice to apply a singular algorithm to the whole country meant that some students were disadvantaged regarding others for being outside of the mainstream framework. This teaches us a bit more about the script of the algorithm, leaning toward a national level comparison to protect the *quasi-market* system but opaque to students' singular trajectories. The application of the algorithm at a national level also allows us to underline which values mattered most to both OfQual and the ministry during the development of the algorithm, being subsequently inscribed in the system. Indeed, by implementing the algorithm at a national level and thus diminishing students' individual trajectories,

⁶Those are privately funded centres, not to be mingled with private students doing their education independently.

it meant that the general standardisation of grade distribution primed on the evaluation of students' abilities. In terms of fairness, this reinforces the idea that OfQual and the ministry believed that it lies in the comparability of students through time. But this also teach us that for them, fairness lies in the applicability of the system to a maximum of students.

The fifth and last trap identified by Selbst et al. (2019) is *framing*, where developers choose a specific representation of reality through the selection of data and labelling. Such selection in turn leads to a lack of understanding of the system in which an algorithm is implemented. To avoid this, Selbst et al. (2019), following on sociologist John Law's argument (Law, 1987), encourage developers to adopt an *heterogenous engineering* approach so that developers can consider the diversity of interactions between a technology and its societal context. Here, maybe due to the various limitations imposed by the short timeframe and the need to deliver grades through the exam boards, OfQual failed to consider the multiple accounts of fairness identified earlier.

Yet, OfQual did try to some extent to implement an approach to identify what mattered to students and teachers through an online survey. But the relatively small number of students amongst the respondents (1,939 students amongst 12,623 respondents) and the general low response rate in comparison to the total number of students being graded (roughly 250,000 students) suggest that it was not very efficient. Moreover, the contestation that followed the announcement of grades also seems to indicate that there was a lack in the inclusion of students and teachers alike in the development of the algorithm. We can identify another account of fairness here where fairness lies in including students in with every aspect of what they are, not only their grades. But most importantly, we can see how the script of the algorithm was contested by the students.

Finally, we can return to Akrich's (1987) argument that as technological objects stabilise and become black boxes, they become instruments of knowledge. In the case of the OfQual algorithm, the accounts of fairness inscribed in the system by its producers (OfQual and the Ministry of Education) conflicted with what fairness meant to its users (students and teachers). This led to a situation where the script of the algorithm was contested, and the stabilisation of the model was challenged. Contesting the algorithm made visible the various negotiations that took place to produce students' grades, highlighting the importance of including students in the evaluation process for them to be able to fight against pre-defined distinction between themselves. This meant that while the OfQual algorithm had all the requirements to become a black box, it remained in a guttered state due to the contestation that arose. Had students and teachers been less averse to the algorithm, it could have stabilised into a black box, producing knowledge that would not be questioned.

5. Conclusion

Throughout this paper I have presented conflicting accounts of fairness that were present in the case of the OfQual algorithm. First, OfQual, the boards and the Department of Education had different views of what fairness meant in the context of the 2020 assessments, although they agreed that their visions were not conflicting. For the Department, the most important thing was that grades were consistent over time, so fairness lay in the comparability of the system over time. For OfQual and the exam boards, I found that fairness lay in the equal treatment of every student assessed in 2020. However, as OfQual developed their system, they had to consider its impact on equity. To do this, they shifted their focus from equal treatment to avoiding discriminatory bias against some protected categories that they could measure. Finally, the demonstration by students and teachers after the grades were announced showed that what they saw as fair was mainly their right to self-determination through assessment.

Throughout the analysis I explored how the production of the OfQual algorithm was entangled in its social context. Unpacking the full assemblage of the algorithm showed how it was produced with the aim of protecting the *semi-market* approach to assessment in England. This allowed me to highlight how the production of the algorithm was shaped by its environment, which left the developers

of OfQual with limited room for manoeuvre. However, because the algorithm was inscribed with the Department for Education's objectives, its script reflected their perspective on fairness. This led to a situation where the algorithm's script was not in line with what its users were expected to do, which was to advocate for a fairer society overall. As I argued in that article, this conflict between the script of the algorithm and the users prevented the system from stabilising and becoming what could be called a black box. Consequently, the protests that followed the distribution of grades underlined the difficulty of making such a system fair, but also the difficulty for a silenced group (the students) to oppose such a system.

Tracing the creation and implementation of the algorithm allowed me to trace how the different negotiations that took place in its creation privileged one perspective of what is fair over others. It allowed us to see how the shifting of the mandatory pass point removed students as a group from the assessment process. This meant that the implementation of the algorithm displaced the accountability of teachers in the assessment process into an unquestionable black box. In doing so, the implementation of the algorithm not only failed to address existing inequalities in the education system, but also worked to institutionalise them.

Indeed, the OfQual algorithm is a clear example of an algorithm that was judged to be fair according to a specific framework set by its makers. However, when looking at the bigger picture, it became clear that the algorithm also implemented some aspects that could be considered unfair by some. In such a context, the concept of biased algorithms should be reconsidered, as the discriminatory results are not the consequence of their code per se, but rather of the societal context in which they are implemented. The question then should not be whether algorithms can be inherently fair, but what account of fairness is promoted when such systems are implemented in contexts that thrive on inequality.

Competing interests. The author(s) declare none.

References

- Akrich, M. (1987). Comment decire les objets techniques? *Techniques Et Culture*, 9, 49–64.
- Akrich, M., Callon, M., & Latour, B. Eds. 2006. *Sociologie de la traduction: Textes fondateurs*. Presses des Mines. <https://doi.org/10.4000/books.pressesmines.1181>
- Benbouzid, B. (2022). L'équité dans la machine ou comment le machine learning devient scientifique en tournant le dos au réalisme métrologique. *Statistique Et Société*, 10 | 3, 69–84. <https://doi.org/10.4000/statsoc.583>
- Benjamin, G. (2022). #FuckTheAlgorithm: Algorithmic imaginaries and political resistance. In 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul Republic of Korea: ACM 46–57.
- Bijker, W. E. (1995). *Of bicycles, bakelites, and bulbs: Toward a theory of sociotechnical change*. MIT Press.
- Bloor, D. (1976). *Knowledge and social imagery* (1st ed.). University of Chicago Press.
- Callon, M. (1984). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. *The Sociological Review*, 32(1_suppl), 196–233. <https://doi.org/10.1111/j.1467-954X.1984.tb00113.x>
- Christin, A. (2020). The ethnographer and the algorithm: Beyond the black box. *Theory and Society*, 49(5–6), 897–918. <https://doi.org/10.1007/s11186-020-09411-3>
- Heaton, D., Nichele, E., Clos, J., & Fischer, J. E. (2023). “The algorithm will screw you”: Blame, social actors and the 2020 A Level results algorithm on Twitter. *PLoS One*, 18(7), e0288662. <https://doi.org/10.1371/journal.pone.0288662>
- Hussain, D. (2020, August 17). Eton College head leads calls to scrap ‘unfair’ A-level algorithm. Mail Online Retrieved August 28, 2024 from <https://www.dailymail.co.uk/news/article-8635439/Eton-College-headmaster-leads-calls-government-scrap-unfair-level-algorithm.html>
- Jaton, F., & Bowker, G. C. (2020). *The constitution of algorithms: Ground-truthing, programming, formulating*. The MIT Press.
- Kelly, A. (2021). A tale of two algorithms: The appeal and repeal of calculated grades systems in England and Ireland in 2020. *British Educational Research Journal*, 47(3), 725–741. <https://doi.org/10.1002/berj.3705>
- Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1), 14–29. <https://doi.org/10.1080/1369118X.2016.1154087>
- Kling, R. (1991). Computerization and social transformations. *Science, Technology, & Human Values*, 16(3), 342–367.
- Laurent, B. (2024). Innovation monsters in the sandbox. *Practicing Anthropology*, 1–8. <https://doi.org/10.1080/08884552.2024.2308437>

- Law, J.** 1987. Technology and heterogeneous engineering: The case of Portuguese expansion. In W. E. Bijker, T. P. Hughes & T. Pinch, (Eds.), *The social construction of technological systems: New directions in the sociology and history of technology* (pp. 111–134). MIT Press.
- Magowan, L.** (2023). Centre assessment grades in 2020: A natural experiment for investigating bias in teacher judgements. *Journal of Computational Social Science*, 6(2), 609–653. <https://doi.org/10.1007/s42001-023-00206-x>
- OfQual.** (2020a). Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: Interim report (No. 20/6656/1; p. 319). OfQual.
- OfQual.** (2020b). *Equality impact assessment: Literature review* (RESEARCH AND ANALYSIS No. Ofqual/20/6611). Retrieved March 12, 2022 from https://assets.publishing.service.gov.uk/media/5e971f1de90e071a145ec51f/Equality_impact_assessment_literature_review_15_April_2020.pdf
- OfQual.** (2020c). *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* (Consultation No. Ofqual/20/6614). Retrieved May 2, 2022 from https://assets.publishing.service.gov.uk/media/5e972a8986650c2dcac1691a/Exceptional_arrangements_for_exam_grading_and_assessment_in_2020.pdf
- OfQual.** (2020d). *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* (ANALYSIS OF CONSULTATION RESPONSES No. Ofqual/20/6614/1). Retrieved March 12, 2022 from https://assets.publishing.service.gov.uk/media/5ec650d486650c76b02b2cca/Analysis_of_consultation_responses_21MAY2020.pdf
- Opposs, D., Baird, J.-A., Chankseliani, M., Stobart, G., Kaushik, A., McManus, H., & Johnson, D.** (2020). Governance structure and standard setting in educational assessment. *Assessment in Education: Principles, Policy & Practice*, 27(2), 192–214. <https://doi.org/10.1080/0969594X.2020.1730766>
- Ozga, J., Baird, J.-A., Saville, L., Arnott, M., & Hell, N.** (2023). Knowledge, expertise and policy in the examinations crisis in England. *Oxford Review of Education*, 49(6), 713–731. <https://doi.org/10.1080/03054985.2022.2158071>
- Reshuffle** 2021: Boris Johnson fires Gavin Williamson as he rings cabinet changes. (June 24 of 2022). BBC News. <https://www.bbc.com/news/uk-politics-58571935>
- Seaver, N.** (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data and Society*, 4(2), 205395171773810. <https://doi.org/10.1177/2053951717738104>
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J.** (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. <https://doi.org/10.1145/3287560.3287598>
- Smallman, M.** (2022). Multi scale ethics—why we need to consider the ethics of AI in healthcare at different scales. *Science and Engineering Ethics*, 28(6), 63. <https://doi.org/10.1007/s11948-022-00396-z>
- Smith, L., Hong, & Hart.** (2020). Written evidence submitted by Smith, Low, Hong and Hart. Retrieved March 7, 2025 from (Written Evidence No. CIE0322; The Impact of COVID-19 on Education and Children's Services). <https://committees.parliament.uk/writtenevidence/8564/pdf>
- The impact of COVID-19 on education and children's services—Committees—UK Parliament.* Retrieved March 7, 2025 from. <https://committees.parliament.uk/work/202/the-impact-of-covid19-on-education-and-childrens-services/publications/2020>
- Williamson, G.** (2020). *DIRECTION UNDER S 129(6) OF THE APPRENTICESHIPS, SKILLS, CHILDREN AND LEARNING ACT 2009*. Retrieved June 13, 2022 from 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
- Winner, L.** (1986). *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press.
- Winner, L.** (1993). Upon opening the black box and finding it empty: Social constructivism and the philosophy of technology. *Science, Technology, & Human Values*, 18(3), 362–378.