Mapping the OfQual algorithm

A socio-technical analysis of an algorithm used to grade students

Merlin Tieleman

m.tieleman@student.maastrichtuniversity.nl

i.6136561

Maastricht University – Faculty of Arts and Social Sciences

European Studies of Society, Science, and Technology

30.06.202

Wordcount:18458

Supervisor:

Pierre Delvenne

pierre.delvenne@uliege.be

Keywords: Algorithm, social construction of technology, mapping, OfQual, discrimination.

Abstract

As algorithms are becoming a more and more important part of our lives, the harms caused by those become more frequent. Understanding how such systems can be biased and the consequences of their implementation within society is thus crucial to prevent further harm. Building on classic STS approaches, we will try to develop and apply new methodologies to the study of algorithms to understand them as part of our societies. First, we will study how to approach algorithms with an STS perspective. Then, we will apply those approaches to the study of the OfQual algorithm used to grade students in England in the summer of 2020. As that algorithm was accused of being discriminatory, its study will allow us to render visible its inner workings and explore the reasons of the resulting fiasco.

Acknowledgements

First, I would like to begin by thanking Mr Jean-Baptiste Fanouillère and Mr Pierre Delvenne who supervised the writing of my thesis. Their continuous support as well as their many comments allowed me to go further in my research than I ever hoped for.

Then, I want to thank my parents and friends who supported me throughout the redaction of this paper. I want to express a special thanks to Sarah, Johan, and Nathan for their encouragement. A special thanks to Clarisse for her support. And a special Thanks to Shannon for everything.

Finally, I want to express my gratitude to all the students from STS, who made this experience unique.

Table of content

INTRODUCTION

In 2018, **Joy Buolamwini** and **Timnit Gebru** released their seminal article: *Gender Shade*. In that paper, they revealed that face recognition algorithms, such as the one sold by IBM and Microsoft, were less accurate on darker-skinned females. That meant that those were at higher risk of being wrongfully mistaken for somebody else, or unidentified. They effectively showed that algorithms were not neutral systems but that they could enforce biases from their makers.

In 2019, Apple was under the spotlight as a whistle-blower revealed that their credit cards' admission program was biased toward men (**Knight**, 2019). Even with a higher credit score, women were more likely to see their credit limit lowered. While the algorithm was not trained on the variable of sex or gender, it had learned to differentiate between men and women. Once again, that showed the consequences that such algorithms can have on their users, whether they are voluntarily exposed to the algorithm or not.

In 2021, Facebook had to publicly apologise after their algorithm asked a user if he wanted to "keep seeing videos about primates" (**Mac**, 2021) after they watched a video featuring a black man, with no primates to be seen. Facebook explained that they recently tinkered with their algorithm and that they still have more progress to make, without going into the details of what could have been the source of the problem (**Mac**, 2021). Nonetheless, their algorithm had produced a harmful prompt that offended many.

A lot of other examples of biased and discriminatory algorithm exist, from racially biased predictive policing (**Christin**, 2020) to unfair Dutch's unfair tax fraud identification software (**Heikkilä**, 2022), and many more will come. Indeed, as algorithms are becoming a more and more important part of our lives, the limitations of such systems are becoming more and more apparent. Understanding how these systems can be biased and the consequences of their implementation within society is thus crucial to prevent further harm.

Engaging critically with algorithms will allow us to better comprehend their limitations as well as their advantages. But studying algorithms in relation to society is not an easy task. While science and technology studies (STS) have proposed numerous ways to engage with artefacts and their environment, the multiple interactions engaged by algorithms as well as their digital format have raised new obstacles for such studies.

Adapting classic STS approaches to the study of algorithms, we will thus try to develop and apply new methodologies to the study of algorithms. To do so, we will divide this work into two main parts. First, we will study how to approach algorithms with an STS perspective. Then we will apply those approaches to the study of an algorithm used to grade students in England in the summer of 2020.

During the first part, we will first review the classic STS theories relevant to the study of artefacts as embedded in their social context. Then, we will try to study what counts as algorithms and what are the limitations defining them. After, we will identify how researchers from different backgrounds engaged critically with algorithms and how contemporary STS scholars have implemented these approaches to their own study. Next, we will consider how different researchers have built on those approaches to render biases within the making of algorithms visible. We will conclude by summarising those approaches into a methodological framework that we can apply to our own case.

During the second part of this thesis, we will apply our framework to the algorithm developed by the English office of qualifications (OfQual) in order to cope with the cancellation of exams. We will start by identifying its place in the English education system as a whole. Then, we will study why the algorithm was put in place. We will identify how the black box that is the algorithm came to be and the different negotiations that took place in its development. We will then study the consequences of the implementation of such a system on the English educational system in 2020. Finally, we will try to render visible the system put in place by OfQual. But first, as the OfQual algorithm will be the backbone of our study, we will start by a brief summary of what happened.

PART 1: THEORY

The story of the OfQual algorithm

In response to the covid-19 outbreak in England in 2020, it was announced that in-person education would be suspended. The announcement was made by the then secretary of state for education, Gavin Williamson, and it meant that exams would thus not take place. As we will later see in detail, **Williamson** (2020) tasked the office of qualification (OfQual) with coming up with a way of providing students with grades that they would have had if exams had not been cancelled.

OfQual is a non-ministerial governmental organisation, which depends on the Department for Education. It is the organism in charge of ensuring that examinations are comparable across subjects and across time. They are in charge of ensuring the due process of examinations to make sure that all students are assessed following fair rules. To do so, they commission the exam boards, who are tasked with delivering examinations and education certificates.

The exam boards are private third parties organisations that mark and award general certificates of secondary education (GCSEs) and advanced levels (A-Levels). According to **Opposs** (2020), boards are typical of a semi-market system in which examinations are made following a logic of competition. That lies on the belief that as boards each want to have the highest achieving students, they uplift the general level of education through competition between them.

As comparability is a major aspect of the English education system (**Newton et al.**, 2008), the Department for Education felt that it was important to ensure the quality of the qualification in 2020. Their main concern was that the cancellation of exams would lead to a greater proportion of students being awarded better grades, thus leading to disparities in the standards set through the years.

That fear came from the year-on-year prediction about students' success. Indeed, schools are asked each year to provide centre-assessed grades (CAGs) for superior education to anticipate the number of students coming their way. Those are given by teachers as they try to guess the final grade of their pupils. But, as we will later see, those are often higher than the actual results of examinations. The Department for Education thus feared that without the examinations, the standards of their education system would be disturbed.

The exams concerned were the A-levels and the General Certificate of Secondary Education (GCSE) as they are key exams in the English educational system that students have to progress[1]. A-Level are passed around 18 years old to get into higher education, the higher the score on A-levels, the more chance a student gets to be accepted in a desired university. On the other hand, CGSEs are used two years prior to access to A-Level preparation. We will mainly focus our interest on the A-level as these can deny the access to a top-university if a student fails them.

In wake of these particular circumstances and the objectives of the Department for Education, it was thus asked of OfQual that grades followed a profile similar to previous years in order to diminish the grade inflation that could happen if the students were graded solely on the appreciation of their teachers (**Kelly**, 2021). Taking into consideration the short time period at hand, OfQual built a statistical model that could be applied to each student through exam boards in order to keep in line with the standards of previous years (**OfQual**, 2020c).

The project was accepted, developed, and put in place in England. But as the students were receiving their grades around the 13th of august 2020, contestation started to arise. Students were accusing the government of putting their fate in the hand of an algorithm that wronged them (**Hao**, 2020). The hashtag *#fuckthealgorithm* was also used on social media platforms such as twitter to build contestation against the algorithm (**Benjamin**, 2022) as it appeared that the algorithm may have downgraded the grades of many students (**Kelly**, 2021).

Indeed, with the use of the model, 40% of students ended up with lower grades than what their teacher had predicted. That meant that in some situation, the teachers had predicted that a student was going to achieve the grade A* (the highest), and the student ended up with a B. We will later study the technicalities of the algorithm and the different reasons that have led to this downgrading of students. But for now, it is important to note that this general contestation led to a cancellation of the calculated results by the authorities (**Kelly**, 2021).

[1] Those are the two main national qualifications that are required for students to move forward. Other levels of education do not have similar examinations.

Studying artefacts with STS

Before approaching the black boxes of algorithms, a detour by some classics of the STS literature is needed. This will allow us to get a better understanding of the general approach used throughout this work. As we will see, studying an artefact in relation to its social environment is not always straightforward. Rooting our work in such literature is thus an important step to understand our general direction.

Opening black-boxes is a prominent approach amongst scholars of science and technology studies (STS). But what counts as a black box? In a seminal article, **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" (p.365). They are placeholders for complex and opaque systems.

That article was based on his previous works on opening black boxes. In his paper about the politics of artefacts, **Winner** (1986) opens the black boxes of technologies such as Moses' bridge, a bridge built by an architect called Moses that kept poorer people from accessing some areas (p.23), or mill-machines enforcing capitalist structures (p.30). His study revealed how those are imbued with explicit or implicit political purposes. With his analysis, thus showed the power relation at play around certain artefacts. Doing so, he distinguished two main ways into which artefacts have politics.

First, with artefacts such as Moses' bridge (p.25), he identified how certain technologies are "stacked in advance to favour certain social interests" (p.26) and that, as a consequence, some people are "bound to receive a better hand than others" (p.26). Some artefacts are thus imbued with politics beforehand in order to put these politics in place. Building on this approach of such technologies, we should thus pay attention to how some technologies provide ways of establishing power.

The second set of artefacts studied by **Winner** (1986) are those with inherent political incentives. Following on the work of **Engels** (1872), he argues that cotton-spinning mills are imbued with capitalists incentives they set the rhythm for the workers, thus pushing them to produce more. According to **Winner** (1986), other large-scale artefacts such as nuclear power plants or pipelines also come with political incentives in order to operate them. He also argues that artefacts might offer a degree of flexibility in regard to those incentives. Some might then be less binding than others.

Building on such work, it is thus important to be careful both to how a technology was imbued with particular objectives, but also how it offers some kind of flexibility around the behaviours enforced by the artefact. As we will later see, STS scholars engaging with algorithms ethnographically (**Seaver**, 2017; **Christin**, 2020) have paid a great deal of attention to those aspects within their field.

Another reference in STS literature in regard to opening black boxes, is the work made by **Bijker** (1995). Amongst other things, he studied the making of the safety bicycle and showed how the making of an artefact was negotiated within a societal context, which came with its influence and limitations. He argued that the identification of relevant social groups was important to understand the development of a technology. According to him, identifying those allows us to understand the negotiations going on in the development of a technology

By identifying such groups, he proposed to realise what he calls the *social construction of technology* (SCOT) approach. That approach encourages researchers to do a sociological deconstruction of an artefact in order to identify its different interpretations by different social groups. By doing so, it allows us to study how the convergence of those interpretations lead to the stabilisation of a new technology as a commonly accepted artefact.

To study the case of OfQual with a SCOT approach would thus mean paying particular attention to how the different groups relevant to its implementation had different interpretations of the artefact. These differences in interpretations lead to a near closure of the technology that was destabilised by the students' protest afterwards due to a different interpretation of the artefact.

Lastly, the *actor network theory* (ANT) developed by **Latour**, **Callon** and **Akrich** (2006) also offers tools to open black boxes. Their main objective was to suggest new types of interactions between the dichotomies of nature and society. One of the key points of their approach being the study of the *sociology of translation*. The study of *translation* allows them to identify how different actants - human and non-human – changed over time.

In order to open black boxes, they studied those as *assemblages* of humans and non-humans. The term is a reference to the work of philosophers **Deleuze** and **Guattari** (1980). It suggests that different actants entre in relations during a period of time. During that time, the actants are intertwined and they become an assemblage that is the result of their relations. We can thus identify those actants and the relations taking place between them.

With ANT researchers should thus study black boxes as a process of *enrolment* between the different actants which result in an *assemblage*. Studying the breeding of scallops, **Callon** (1986) thus studied how actants negotiate together and achieve their objectives by *interesting* others. Doing so they create an obligatory passage point (OPP) through which all objectives. The identification of said OPP and the way the actants are translated through it is consequently a major part of an ANT analysis.

Classic STS literature thus offers different approaches to study the relation between society and technology. Some of these theories offer great tools to open black boxes in practice. We will later see how these theories apply to the study of algorithms and how we might build on these tools to study them. We will also see some limitations of said theories to engage with algorithms critically. But in order to do so, we first have to understand what is at play when we study algorithms.

Studying algorithms

Discussing algorithms is not as straightforward as what may seem. Indeed, the term algorithm is used loosely by many actors, whether they are producing algorithms or studying them. Moreover, algorithms are often replaced by other terms such as artificial intelligence or machine learning. According to **Seaver** (2017), many practices of algorithms exist and defining algorithms is thus a crucial step to engage with them.

We might start with a generally admitted, technical definition, that we can find in a dictionary. Of course, such a definition is still imbued with a particular understanding of algorithms. It is the result of a "deeply interpretative, political process" (**Joyce** et al., 2021, p.2) as the dictionary tries to put boundaries around the notion of algorithms. Yet, regardless of those limitations, it gives us a point of entry to start our analysis.

According to the **Oxford English dictionary**, algorithms in computing can be defined as "a procedure or set of rules used in calculation and problem-solving; […] a precisely defined set of mathematical or logical operations for the performance of a particular task". That first definition checks-out with studies about how algorithms' makers define them. Indeed, in a study on the discourse on algorithms, English computer scientist **Dourish** (2016) underlines that professionals define algorithms as “an abstract, formalized description of a computational procedure” (p.3).

Moreover, it is not sufficient to settle for a technical definition. Indeed, it would make us blind to many social aspects of the making of algorithms that are put aside by such an approach. As argued by French sociologist **Cardon** (2019), algorithms contain “the principles, interests and values of their creators: the operational implementation of those values go through technical choices, statistical variables, chosen thresholds and calculation methods”[2] (p.356). That definition can also be found within a book by the engineer **Panos Louridas** (2020). He argues that “programming [algorithms] is the discipline of translating our intentions to some notation that a computer is able to understand” (p.22).

We thus now are confronted with the idea that algorithms are not merely technical. Indeed, according to **Cardon** (2019) and **Louridas** (2020), those are more socio-technical translations of human agency. In a similar vein, the socio-anthropologist **Joyce** et al. (2021) define artificial intelligence as being “about the deployment of computing infrastructure and programming code to create systems expected to mimic, augment, or displace human agency” (p.2). But already, a semantic change occurs. We went from defining algorithms to an author talking about artificial intelligence (AI). But what are the differences between both if there are any?

According to **Panch** (2019) and his team studying algorithms and AI, the latter is defined as “a family of techniques where algorithms uncover or learn associations of predictive power from data” (**Panch** et al., 2019, p.1). According to them, AI is an improvement on algorithms through different techniques. In another paper, they identify five of these techniques[3], the most prominent being machine learning where algorithms learn patterns in a large dataset (**Panch** et al., 2018). We may thus say that AI is an improved – or complexified – form of algorithm. With that said, when discussing AIs, we are in fact talking about algorithms. Nonetheless, all algorithms are not AIs.

[2] Translated from French by me.

[3] Those being: machine learning, deep learning, supervised and unsupervised learning and reinforcement learning (Panch et al., 2018).

To summarise **Panch** (2019) perspective, algorithms are thus complex systems which build on techniques to acquire a form of agency. There is therefore a variation in complexity between the different concepts we saw. Algorithms are a complexified form of code, and they can in turn be complexified to become AI. It is important to be aware of those different degrees of complexity around a common code in order to better understand what is at stake when studying algorithms and where their complexity comes from.

We may thus say that algorithms are opaque systems that we might only analyse in terms of what comes in and out. That approach indeed prevails amongst some studying algorithms critically (O'Niel, 2016; Buolamwini & Gebru, 2018), but it leads to the idea that algorithms can only be studied in terms of input and output. That approach should remind us of the study of black boxes in STS. Indeed, we saw earlier that classic STS scholars (Bijker; Winner; Latour; Callon; Akrich) developed tools to open black boxes (SCOT, politics of artefacts, ANT). Some of their approaches can now be used in order to build a comprehensive approach to algorithms.

Doing so, STS scholar **Angèle Christin** (2020, p.3) identified four main components that lead to the black boxing of such systems. First, their *intentional secrecy* which is imbued by the one making them. That means that companies making algorithms tend to keep their code secret in order to protect their intellectual property. Getting access to the code of the algorithm might be difficult. Second, the *technical illiteracy* surrounding them. Due to their complex working and the skill level needed to study them, most people are unable to understand the code. Third, the *untelligibility* of some of the decisions taken by algorithms as those are speaking in their own inapprehensible language of some sort, so that even specialists may not always understand the decision taken by the system. And fourth, their *size* as they work with huge datasets. According to her analysis, those are the reasons why algorithms are seen as black boxes that may only be comprehended in terms of input and outputs.

With that analysis, **Christin** (2020) argues that the opaqueness of algorithms is in fact the result of the relations within an assemblage between code and humans. Following on such approach, digital anthropologist **Nick Seaver** (2017, p.5) 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.

This reminds us of the STS theories we saw previously as it pushes us to see science and technology as being in constant negotiation with society. As the scallops studied by **Callon** (1986), there are no actants apart from the network, all of them are in the same *assemblage* in constant negotiation. This goes in line with the study of socio-technical artefact as being one within a system where human and non-human are in constant interaction.

When following views promoted by **Christin** (2020) and **Seaver** (2017), it becomes difficult to propose an operational definition of algorithms that could be used to study them in general. Each algorithm will be opaque in their own way and have their own specific interactions. In order to build a strong theoretical frame to study our own case, we thus have to see how the algorithms are studied as part of society. In STS of course, but also in other fields to understand how we might be able to engage with them in practice.

Engaging with algorithms critically

Many authors have engaged with algorithms critically. Doing so, those authors underline how the use of algorithms reproduce or exacerbate existing discrimination within society. Building on the work of **Christin** (2020) and **Seaver** (2017), we are going to see how those approaches are in line with the STS approaches we saw earlier. This will allow us to identify ways to engage with algorithms critically and pitfalls to avoid.

Christin (2020) identified how different authors engaged with algorithms critically. We will see how different approaches have been put in place. The first approaches we are going to see have engaged with algorithms in a way that does not allow us to fully understand their entanglement within society. Still, building on the following methods is important as it allows us to better understand the pitfalls we may be facing when engaging with algorithms.

First, some specialists began to build a critical approach to algorithms on the basis of *algorithmic audits* (**Christin**, 2020). These approaches rely "on statistical and computational methods in order to examine the outputs of algorithmic systems, specifically […] their discriminatory impact" (p.3). **O'Neil** (2016), **Buolamwini & Gebru** (2018), and **Birhane** (2021) all built on such an approach to engage critically with algorithms.

Their studies have in common that they analyse algorithms with quantitative tools in order to identify their statistical biases. Of course, such analyses are very useful to render discriminations in algorithmic decisions. But they only take into consideration the algorithm in terms of input and the output. The practice of audits thus builds on the opaqueness of algorithms instead of trying to unpack them. Yet, that approach can be used in relation with the approach promoted by **Seaver** (2017) in order to study how the inputs and outputs of algorithms are entangled within society.

Yet, it is important to note that work is being done by defenders of *algorithmic audits* to push for a greater consideration of the general social context of the algorithms. In a recent paper, **Costanza-Chock et al.** (2022) argue that audits should pay more attention to the structural forces at play around the algorithm. By doing so, they want to stand out from purely quantitative analysis to better account for their biases.

The second set of approaches identified by **Christin** (2020) are the *cultural and historical critique* of algorithms. According to her, these approaches pay a greater attention to the particular epistemology of algorithms in their societal context by acknowledging their complex interactions with their environment. They allow us to underline how algorithms encode dominations embedded in our society. Scholars using these approaches typically identified power asymmetries in information and surveillance. Finally, they identify the broader role of algorithms in a neoliberal system.

Amongst people using this set of approaches, there is the book by sociologist **Ruha Benjamin** (2019), *race after technology*. In that book, she qualifies algorithm of being the *new Jim code*, referring to the Jim Crow laws enforcing racial segregation. She argues that due to their history and implementation, algorithms sustain existing discrimination based on race. Another approach is done by **Steffen Mau** (2019), another sociologist. In *the metric society*, he analyses how self-tracking apps turn bodies in data-driven competitions.

But as these approaches place algorithms in a broader socio-cultural context, they tend to be too generalistic. Indeed, while building macro-level analysis, those authors neglect the role played by the artefact that is the algorithm. They tend to represent it as a mere tool for bigger discriminations. Building an approach between *algorithms audits* and *cultural and historical critique* of algorithms is thus necessary to be able to study both the interactions within a wider societal context and the role played by the algorithms in such interactions.

To do so, **Christin** (2020) identifies a last way to critically engage with algorithms, the *ethnographic studies* of algorithms. According to her, the ethnographic study of such systems allows for a better understanding of algorithms within their socio-cultural context by focusing on two main sites of study. Namely the making of algorithms and their interaction within society. But, as she underlines: "ethnographers can only study places and practices to which they have access" (**Christin**, 2020, p. 7). This of course poses a problem to the study of algorithms as those are opaque due to the different characteristics that we saw earlier.

In order to overcome the difficulties brought by the study of algorithms as black boxes, she encourages scholars to go beyond that perception when studying algorithms (p. 8). To do so, she built on her STS background and proposed to use an approach based on ANT. She argues that algorithms could be studied in terms of *assemblage* and *enrolment*. As she argues, "instead of focusing on algorithmic 'black boxes', ethnographers can study how collectives of human and non-human actors emerge, solidify, and evolve over time" (**Christin**, 2020, p.10).

This goes in line with the approach proposed by **Seaver** (2017) as he proposed to engage with algorithms as "unstable objects, culturally enacted by the practices people use to engage with them" (p.5). Both argue that by studying algorithms as such, scholars could comprehend them as actants within a relation instead of an outsider force.

In a similar vein to those approaches, the social scientist **Rob Kitchin** (2017) suggests "unpacking the full socio-technical assemblage of algorithms" (p.25). What he proposes is to bring to light the complex intertwining of the algorithms through a methodical process of opening it. Doing so, the researchers engage with algorithms as a whole, taking into consideration all the infrastructure surrounding them.

Yet, doing so is no simple task. We have seen how researchers practising *algorithmic audits* used tools to produce quantitative analysis of algorithms. In order to build an *ethnographic approach of algorithms*, **Christin** (2020) proposes three methodological tools. Those are made to engage with algorithms in order to better comprehend their interaction with society.

First, she argues for the study of *algorithmic refraction*. As with refraction within the laws of physics, that approach suggests studying algorithms in terms of prisms that both "reflect and reconfigure social dynamics" (p.11). This approach allows scholars to study algorithms as a chain of human and non-human dynamics and avoid the opaqueness of black boxes. That approach is important for the study of the OfQual algorithm as it invites us to look at how the grades predicted by teachers were refracted by the algorithm.

Then, she proposes to practise *algorithmic comparison*. As it has already been shown in STS (**Jasanoff**, 2009), comparative studies allow to identify the key features of each case studied. By comparing algorithms in their usage, scholars could understand how their inner workings differ or not without having to shed light on their opaqueness. A limitation of such methodology being the need for two algorithms sufficiently similar to compare them.

Finally, she argues for *algorithmic triangulation*. She encourages scholars to embrace the algorithms within their studies. According to her, embracing the use of algorithms might help to deal with many difficulties met by ethnographers. For example, some algorithms might help to tag interviews more efficiently. Also, using algorithms to engage with their fields of study, by engaging with people online for example, could help the researcher to disengage from their field. The term triangulation is used in reference to the incentive, for social sciences, to combine "research methods, angles, and material in the study of the same phenomenon" (**Christin,** 2020, p.12).

Those techniques allow researchers to study algorithms, not as being *in* culture, but *as* culture as suggested by **Seaver** (2017, p.5). Doing so, the algorithms are not machines or tools apart from society that require special techniques to study, but rather part of their socio-cultural system that might be understandable with those approaches. With an ethnographic approach to algorithms, social scientists are thus able to engage with algorithms critically with tools specific to their discipline. This is important as we saw earlier that more technical approaches done by computing specialists may fail to take into consideration the influence of society on the making of algorithms.

Yet, a difficulty remains for social scientists. As we saw with **Christin** (2020), the black box aspect of algorithms is in part due to the technical illiteracy of the layperson. As for other subjects in STS, it is the task of the researcher to build a comprehensive analysis of the object they study on the basis of their technical comprehension. To do so, **Kitchin** (2017) suggests dismantling the code of the algorithm along with the social interaction happening around it. The case of OfQual is a great one to do so as they tried to render the making of the algorithm as open as possible by publishing a lot of documents explaining how it was put in place.

In conclusion, we saw that studying algorithms as isolated artefacts does not allow for a comprehensive analysis of their interactions within a societal context. In the case of OfQual, it would thus not be sufficient to analyse the inputs and outputs of the algorithm as it would not allow us to comprehend how the algorithm was enacted by the different actants around it. Such an analysis would render invisible many of the interactions that happened around the algorithm.

But while we saw ways to open algorithms and engage with some of their components, those approaches do not offer ways to render their entanglement within their context, and the power relations ensuing. Still, we saw that the methodologies proposed by **Christin** (2020) and **Seaver** (2017) are closely related with the notion of *assemblage*, which is a hallmark of STS literature. Following, we will thus see how that entanglement between artefact and society has been rendered. First in STS with **Latour** (1992) and then within the specific field of algorithms.

Mapping algorithms

As we saw, the notion of *assemblage* in STS literature is strongly linked with ANT (**Akrich et al.**, 2006). As with black-boxes in the breeding of scallops (**Callon**, 1984), STS scholars had to unpack the *assemblages* within the black-box. In order to open socio-technical systems and render them visible, **Latour** (1992) proposes to realise socio-technical graphs (STG). Those are maps of the controversies happening around an artefact.

To do so, he proposes to look at the modification undertaken by the *dictum* of the artefact in regard to the *modality*. These are references to linguistics, where the dictum is the essence of the sentence, it cannot be changed without changing the meaning, while the relation to the *dictum* is set by the *modality*. What **Latour** (1992) suggests is that we look at how the artefact change in regard to the modality. How the artefact changes in nature as if its *dictum* was changing. That change is the result of the *translation* undertaken by the artefact.

According to him, STGs would allow us to identify the translation undergone by an individual or a collective. He proposes to map scientific controversies in order to build over the old dichotomy of human and non-human. To do so, he encourages researchers to identify how an artefact evolves over time. He argues that we should both follow how an artefact changes in nature and how it evolves over time, as he suggests with the following map. Studying how an artefact changes in nature and in discourse would then help us to overcome the old dichotomy by paying less attention to whether an actant is human or not.

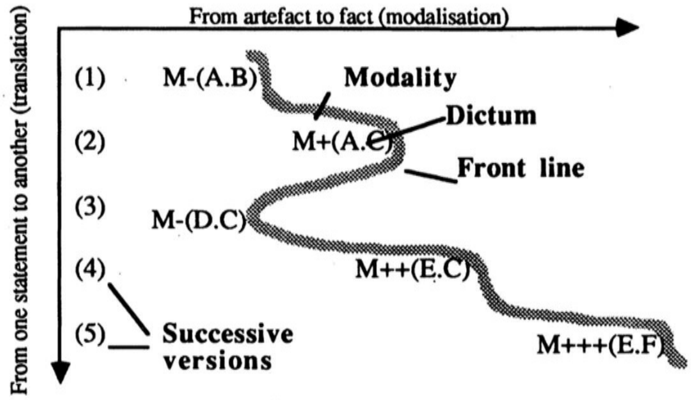


Figure 1: Illustrative socio-technical graph made by Latour.

On this illustrative map, **Latour** (1992) shows how the *dictum*, the nature of the artefact, changes over the successive translations. The (+) and (-) shows the relation of the modality, set by the users. The front line shows how the artefact changes over time with complete rejection of it on the left and complete acceptance on the right. At the end, the *dictum* is completely different from the beginning, suggesting that the artefact had undergone a complete transformation. In the case of algorithms, we can for example follow how code changes in nature in regard to the modality to stabilise as an algorithm.

Rendering visible controversies around a black box, has also been done in the study of algorithms. One clear example of that could be the map proposed by **Crawford** and **Joler** (*Anatomy of an AI System*, 2018) observable at the end of this section. Indeed, in an effort to render visible the inner workings of Amazon's Alexa system, they tried to provide a complete scheme of the system. They started from the extraction of primary resources in mines to the decomposition of the components after its disposing. That map is complemented by a book by **Crawford** (2021) in which she tries to develop a generalist atlas of AI.

The map starts in the bottom left corner, with the extraction of primary resources. It then goes up the chain of production. During that part, we can identify the different actors in the making of the Alexa system. Then we go to the right to see how the system works and how it is used. That shows the different interactions with and within the system. The map then finishes on the right with the disposing of the artefact. From up to bottom, we can see how it goes from trash can to disposing sites, to geological processes.

But, in her book as well as in the map, the training of the algorithm remains quite opaque. Indeed, **Crawford** (2021) argues that there are two types of algorithms at play in machine learning (p.115), the *learning* ones and the *classifying* ones. The first ones are trained with examples and show the others how to analyse *inputs* for a specific *output*. That conception of algorithms as input and output can also be found in the map (Fig.2). Whilst this approach allows us to better comprehend the entanglement of an algorithm within their societal environment, it does not allow us to render how this is enacted within the code of the algorithm.

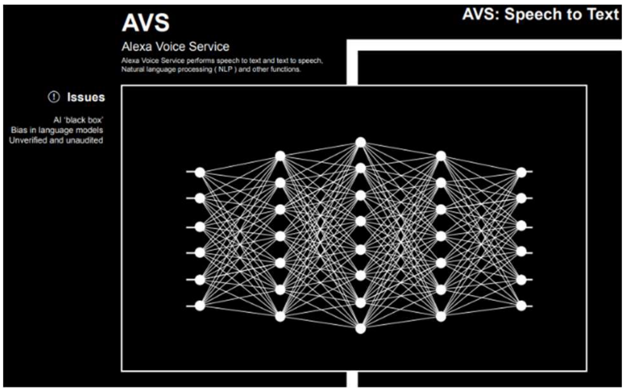

AVS
Alexa Voice Service
Alexa Voice Service performs speech to text and text to speech,
Natural language processing (NLP) and other functions.
AVS: Speech to Text
Issues
AI 'black box'
Bias in language models
Unverified and unaudited

Figure 2: depiction of the training of the AI training as a black box by Crawford and Joler.

In another paper focusing solely on the inner workings of algorithms, science philosopher **Pasquinelli** in team with the aforementioned **Joler** (2020) proposed to open the black box of algorithms in general. They developed a map of the inner workings of algorithms and the different biases faced in the process, also observable at the end of this section. Their map is pretty complete as they considered the many ways in which an algorithm is imbued within a socio-cultural context and the different steps required to build them.

To follow the map, we have to start at the bottom. On the left there are the human biases and interventions while on the right, we have the machine and statistical biases. In the bottom part, they depict how the dataset is made. Then, they show how the algorithm is made and finally how it is applied. According to them, the higher we go in their map, the more bias are amplified, and the information within the algorithm reduced, thus leading to a loss in precision.

But still, some details are missing in regard to our STS framework. Indeed, the lack of connection shown between the different actors obscures the many power relations that exist between them. In reality, it might also be that some of the actants identified in the map are actually the same along different steps. Moreover, they still operate a dichotomy between human and non-human by identifying *machine biases* that reduce the information within the algorithm. Furthermore, while they take into consideration parts of the influence of the societal context on algorithms that aspect is still relatively concise. Such a map built with STS contribution might thus be less straightforward and link the different actors between them.

Finally, an attempt at mapping the OfQual algorithm (Fig.3) was made by the English sociologist **Benjamin** (2022) in order to study the imaginaries of resistance around the algorithm. To do so, they identified all the actants involved in the development and the contest around the algorithm. They then placed them in regard to the level on which they interact. Their goal was not to provide the reader with a complete understanding of the algorithms, it is still rendered as one of the actants that interact with the system as a whole.

To that regard, we saw earlier that **Christin** (2020) identifies two main fields within the *ethnography of algorithms*: the making of algorithms and their interaction within society. The map made by **Benjamin** (2022) is closer to the second field while our research will focus on the first one, the making of algorithms.

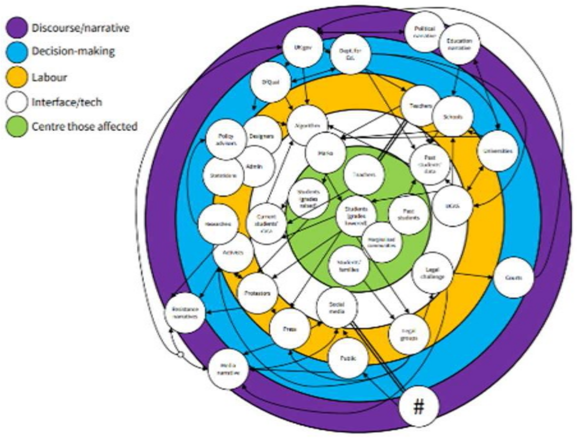
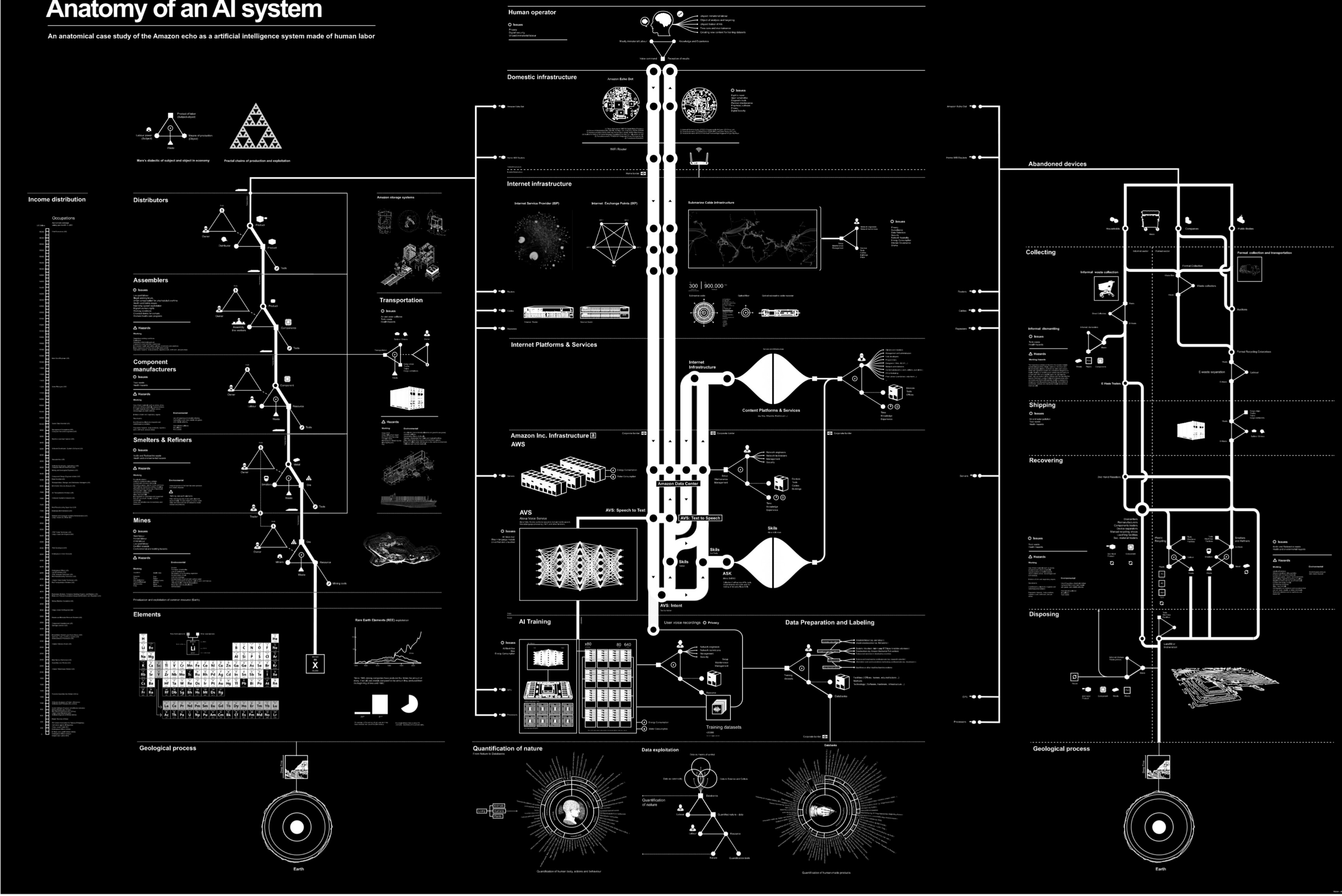


Figure 3: Benjamin's depiction of the OfQual algorithm.

To do our maps, we will thus start by studying how the algorithm came to be and how it evolved over time as we want to build on that development to comprehend how it was enacted by society around it. Yet, as we saw in the other maps, we have to be careful not to fall back in the dichotomy between human and non-human by black boxing the algorithms. Our goal is thus to build on the work of such scholars and on STS contribution to study how the making of the OfQual algorithm was entangled within its societal environment.

Anatomy of an AI system

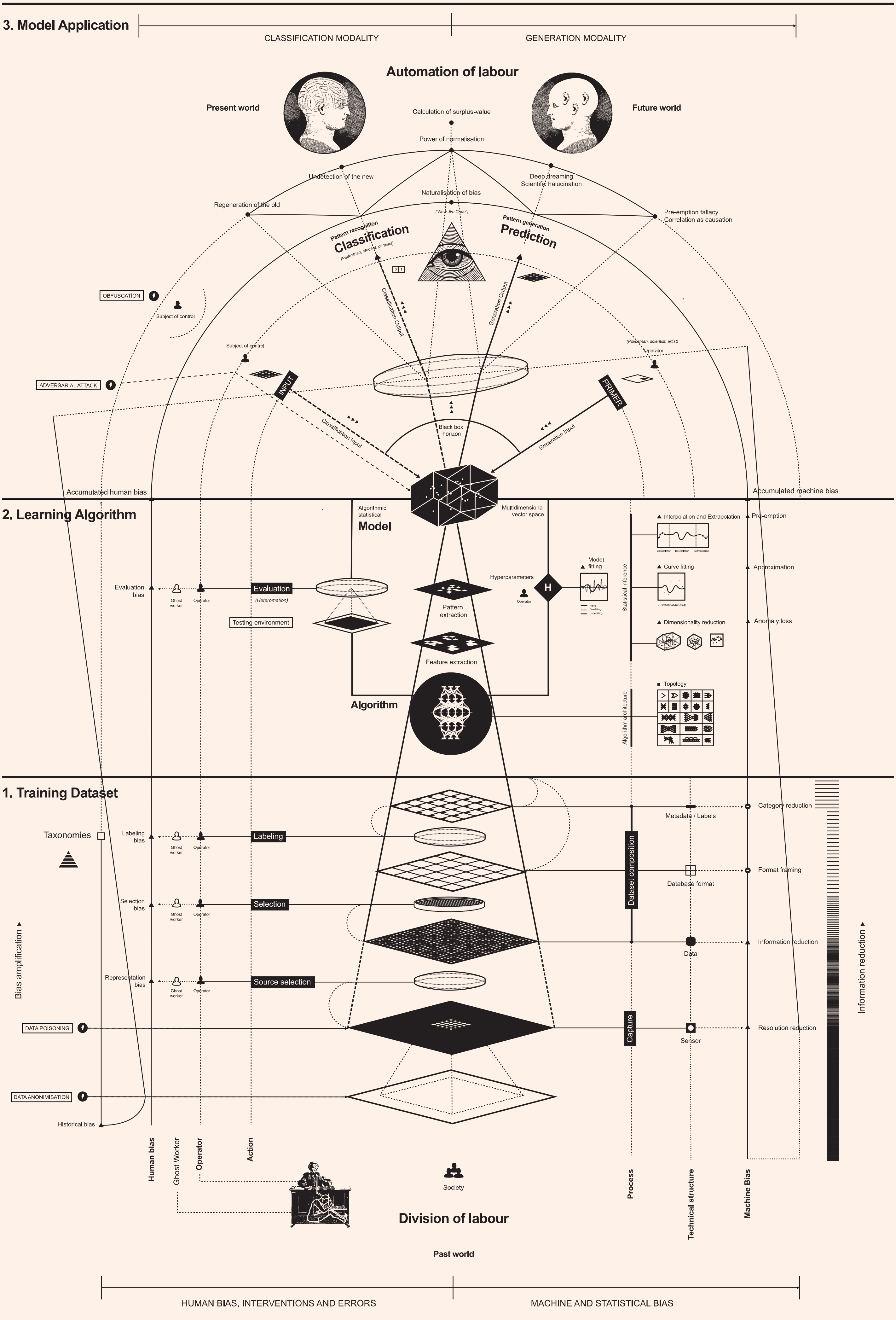
An anatomical case study of the Amazon echo as a artificial intelligence system made of human labor



NOOSCOPE

THE RISE OF AI STATISTICAL MODELS AS INSTRUMENTS OF KNOWLEDGE
A DIAGRAM OF MACHINE LEARNING ERRORS, BIASES AND LIMITATIONS

VLADAN JOLER AND MATTEO PASQUINELLI (2020)
WWW.NOOSCOPE.AI



3. Model Application

CLASSIFICATION MODALITY

GENERATION MODALITY

Automation of labour

Present world

Future world

Calculation of surplus-value

Power of normalisation

Undetection of the new

Deep dreaming
Scientific halucination

Regeneration of the old

Naturalisation of bias

("New Jim Code")

Pre-emption fallacy
Correlation as causation

Pattern recognition
Classification
(Pedestrian, student, criminal)

Pattern generation
Prediction

0 1

Classification Output

Generation Output

OBFUSCATION

Subject of control

ADVERSARIAL ATTACK

Subject of control

(Policeman, scientist, artist)
Operator

INPUT

PRIMER

Classification Input

Generation Input

Black box
horizon

Accumulated human bias

Accumulated machine bias

2. Learning Algorithm

Algorithmic statistical
Model

Multidimensional vector space

Evaluation bias

Ghost worker

Operator

Evaluation

(Heteromation)

Testing environment

Pattern extraction

Feature extraction

Hyperparameters

Operator

H

Model fitting

Fitting
Overfitting
Underfitting

Algorithm

Statistical inference

▲ Interpolation and Extrapolation

Extrapolation Interpolation Extrapolation

▲ Curve fitting

x Statistical Anomaly

▲ Dimensionality reduction

Algorithm architecture

■ Topology

Pre-emption

Approximation

Anomaly loss

1. Training Dataset

Taxonomies

Labeling bias

Ghost worker

Operator

Labeling

Selection bias

Ghost worker

Operator

Selection

Representation bias

Ghost worker

Operator

Source selection

DATA POISONING

DATA ANONIMISATION

Historical bias

Bias amplification ▲

Dataset composition

Capture

Metadata / Labels

Database format

Data

Sensor

Category reduction

Format framing

Information reduction

Resolution reduction

Information reduction ▲

Human bias

Ghost Worker

Operator

Action

Society

Process

Technical structure

Machine Bias

Division of labour

Past world

HUMAN BIAS, INTERVENTIONS AND ERRORS

MACHINE AND STATISTICAL BIAS

Methodology

Engaging with algorithms is a complex task. Some of the approaches being presented here are relatively new and do not always come with fully developed methodologies to study algorithms. Nonetheless, they offer new ways to approach algorithms in regard to their entanglement within society. In order to build our own analysis, we will thus build on their approaches to sees what the OfQual algorithm can teach us.

The theories we have seen so far encourage us to overcome the analysis of algorithms as untouchable black boxes. As **Christin** (2020) identifies how the opaqueness of algorithms is socially enacted, authors such as **Seaver** (2017) and **Kitchin** (2017) propose ways to engage with them in practice. Doing so, they push researchers to dismantle algorithms by paying particular attention to how enrolment happens around and with it.

In order to do so, **Seaver** (2017) suggests practising what he calls *scavenging*. He suggests collecting secondary and sometimes tertiary sources to grasp the algorithm. Doing so, the researchers should not reject resources on the basis of their intelligibility. Indeed, the inaccessibility of sources is also a part of the analysis as **Christin** (2020) argues that it is a major aspect of black boxes. Collecting code and gargantuan datasets is thus also part of such research.

These resources will allow us to dismantle the algorithm as encouraged by **Kitchin** (2017). In practice, he proposes to deconstruct the code step by step, by including all comments and documentation available. Methodologically going through the scavenged resources, we should thus identify the different steps going on within a code to translate an input into an output. And it is through the studies of said translation that we might engage critically with the algorithm to underline its entanglement within society.

In the case of OfQual, as it was part of a public policy in order to overcome the cancellation of exams, they tried to be as transparent as they could in order to build confidence in the system from the public (**OfQual**, 2020c). As a result, a lot of documents were produced explaining the choice made by OfQual and the inner workings of their algorithm. Along with the analysis made by **Benjamin** (2022), this will help us to build a comprehensive analysis of how the algorithm made by OfQual translated students into results.

With those resources, we will retrospectively identify how the algorithm was made in regard to its environment. Following on classic STS authors (**Bijker**, 1995; **Callon**, 1986), we will pay particular attention to how the different negotiations within the implementation of the system led to its contested outcome. Then, as **Winner** (1986) did, we will identify the politics of the OfQual algorithm. Doing so will enable us to underline how it was intertwined with its environment.

Doing so, we will be able to render the process of the algorithm through a mapping of the different translations that took place. But as mapping is essentially a work of description, STS authors such as **Meulemans** and **Tari** (**Seurat** et al., 2021) remind scholars to follow **Bloor's** (1976) *principle of symmetry*. That encourages us to avoid studying a past controversy following a discourse *a posteriori*. We should thus not identify a winner and a loser beforehand due to the outcome of the controversy.

In regard to the case at hand, we should therefore not see the OfQual algorithm as being wrong in advance. We should avoid perceiving it as being inherently discriminatory against some, but we should instead identify how such discourses appeared and where. Doing so will allow us to understand where does the usage of such an algorithm stem from and what are its consequences.

Most of our analysis will thus be built on the grey literature made available by OfQual and others about the algorithm. With those resources, we will rebuild, step by step, the different negotiations that happened while making and implementing the algorithm. To complement those sources, we will expand our analysis with a semi-directive interview of a director of the department within OfQual that developed the algorithm. This work will allow us to decipher each moment when the algorithm was shaped by its environment.

PART 2: ANALYSIS

In the second part of this work, we will apply the theories we have studied to the case of the OfQual algorithm. By doing so, we will try to understand how the different negotiations that took place in its development led to its contested result. To that aim, we will start by identifying the algorithm in regard to the general English educational system. We will then see where the idea of implementing an algorithm came from. After that, we will study the different steps in the making of the system and the consequences of the choices made. Then, we will study the consequences of the implementation of the algorithm in the educational system. To conclude, we will try to render visible the whole case through the process of mapping that we studied beforehand.

The quasi-market system of English education

As we saw at the beginning of this thesis, the English education system could be qualified of a quasi-market system **Opposs** (2020). In order to better understand what it means, we will see how the system works as a whole. By the same occasion, we will identify the different actants that are taking part in the system as well as their objectives.

At the bottom line of the English educational system, the first set of individuals we identify are the students who are being marked. In our case, those are all students passing their A-Levels in England in 2020. Their main objective is to succeed in their exams in order to move on with their lives. Those might be in private or public schools, or even in remote education. According to OfQual (2020c, p.11), they were around 250,000 in summer 2020.

The students have to choose their school on the basis of the A-Levels they want to pass. They chose those in function of what they want to do after. Some universities also require that students pass some precise A-Levels in order to be accepted in their programs. Typically, a student chooses between two and three subjects, leading to about 700,000 A-Levels being delivered in 2020.

Next, there are the teachers. They are tasked with instructing students in said subject and evaluating them through the years. Their goal is to prepare students for their final examinations. Those teachers are part of schools and colleges that are referred to by OfQual as centres (2020c, p.11). It refers to all institutions within which education happens, such as schools and colleges. A student that wants to pass an A-Level will thus go to a centre that provides A-Levels in the subject that he wants but also on the basis of the exam boards to which the centre refers.

Indeed, centres ask of exams boards to provide them with certified A-Level examinations. Each A-Level examination is developed and marked by one of the four exam boards[4]. Some boards offer A-Level in the same subject, the centres are thus able to choose between them. According to OfQual (2020c, p.83), centres might choose a board over another on the basis of the form of the evaluation, the supporting service, or the specifications of each centre.

According to **Opposs** (2020) it is the boards that make it a quasi-market system. According to him, the delegation of assessment to private third-parties organisations lies on the belief that competition between the different boards will push them to raise the level of education overall, as each board wants to have the best achieving students. The system has been contested in England by politicians from different backgrounds, but it is still in application to this day (**Opposs,** 2020).

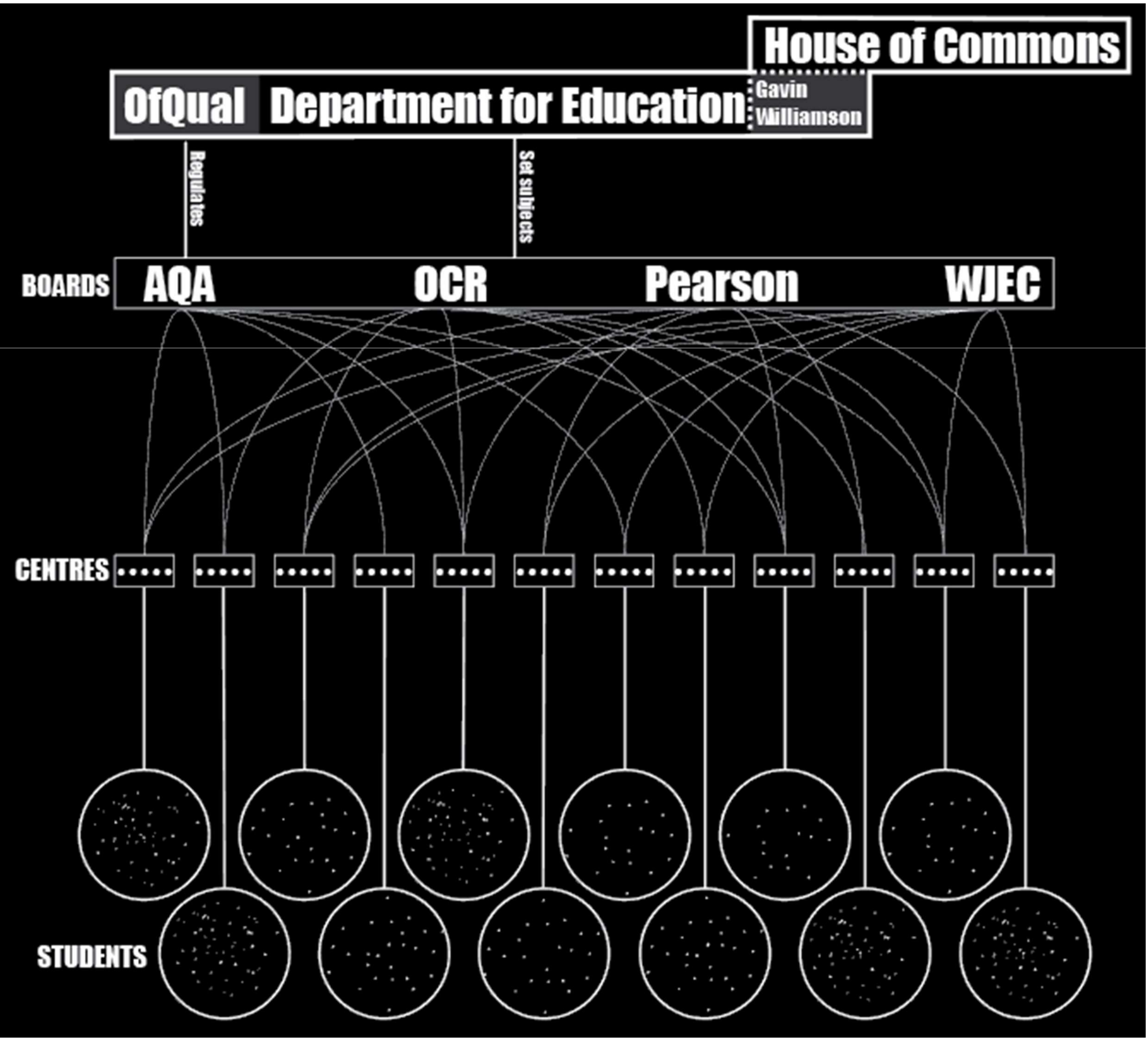
The list of subjects that exam boards are allowed to award is set by OfQual on the basis of content set by the Department for Education. Finally, each exam board provides the examination asked and ensures the correct correction of said evaluations. They then provide students with their grades through their centres via a standardised national platform. The main objective of each board is to have the higher-ranking students in order to have more centres referring to them.

OfQual is specific to England, and it is the only office of regulation for education under the direct authority of the UK government. Indeed, parliaments of Scotland, Wales and Ireland are responsible for their respective matters in education. This is a consequence of the particularities of the quasi-market approach promoted in England as the other parliaments deliver themselves the qualification without referring to third-party (**Opposs,** 2020).

OfQual is divided between different departments depending on their different responsibilities. Its main objective is to regulate the different exam boards to ensure comparability. It is a non-ministerial governmental organisation, which means that they have to follow recommendations from the Department for Education, which is a ministerial organisation, but they might do so with a degree of freedom.

[4] Those are as follow: AQA, OCR, Pearson and WJEC

Above all, there is the government. The content that should be covered by each subject is set by the Department for Education, which depends on the Secretary of State for Education, Gavin Williamson at the time. Finally, Gavin Williamson answer to the House of Commons as they raise questions. The general system of education in a year with examination could be summarised as such:



This graph shows that students are being instructed by centres, composed of teachers. Centres then choose which board they want to refer to for each A-Level. The boards provide examinations and grade the students. The way the boards act is regulated by OfQual, and the composition of each subject is set by the Department for Education. The department depends on the Secretary of State for Education, who answers to the House of Commons.

Building on **Bijker**'s (1995) SCOT approach, that map helps us to identify the relevant social groups that are at play within the English educational system: students, centres, boards, OfQual, and the government. That will be important for the rest of our analysis as the division of those groups into producers and users will lead to specific negotiations during the implementation of the algorithm.

Moreover, looking at the system with an ANT perspective, we already know the objectives of each of those groups. As **Callon** (1986) did with the scallops, we can thus identify an *obligatory passage point* (OPP) through which all of the objectives converge. As we have seen, the students want their grades to move on but OfQual, in line with the government, wants to ensure a degree of stabilisation to keep in line with their objective of comparability. It is the exams boards that successfully *interest* those objectives into the process of examinations.

Exams, delivered by the boards, are thus OPP within the English education system. But as Gavin Williamson announced that schools were to close due to the covid-19 crisis, it was not possible to go through that OPP anymore as it was not allowed to sit exams. But as we will now see, the Department for Education still needed a way to deliver grades in a way that would fit the exam boards and ensure the comparability between the years.

Maintaining the education system at all costs

On the 20th of March, it was asked that students were provided with grades so that they could continue with their lives as smoothly as possible (*Educational Settings - Hansard - UK Parliament*, 2020). In his article about the algorithm, **Kelly** (2021, p.12) argues that it was surprising that universities did not directly use the actual grades of students, through continuous evaluation up to this point, for their applications. This was actually the method used in countries such as France and Belgium. But as we saw, the English educational system relies on exam boards to assess students and deliver certificates, which means that they could not bypass them as they have to deliver a general certificate for each student.

Moreover, the Department for Education asked OfQual "that qualification standards are maintained, and the distribution of grades follows a similar profile to that in previous years" (**Williamson**, 2020). That meant that OfQual had to avoid producing grades that would generally be too high or too small compared to the year-on-year average. According to the OfQual director that I interviewed, that fear of inflation was based on the idea that it is preferable for the education system as a whole to have some sort of coherence in order to be able to compare students in-between each year (**interview**, R12). OfQual was thus tasked with coming up with a way to standardise results for A-Levels and CGSEs as those are the two main qualifications that are eliminatory to access further education.

The main argument to put in place a standardisation process was therefore to follow the year-on-year means of grades. That was made in order to have consistency in grade delivery over the years so that they would be comparable. According to **Newton et al**. (2008), this is an important aspect of the English educational system. It is strongly linked with the logic of competition induced by the semi-market system (**Opposs**, 2020). Indeed, the educational system relies on the idea that grades should be comparable between subjects and time. That meant that obtaining the highest grade in 2020 should not be harder nor easier than it was during previous years.

Whilst **Newton et al.** (2008) recognise that the comparability of the system is a subject to political debates, "filling many newspaper column-inches each year" (p.10), there were no such debates when the cancellation of exams was announced. Indeed, looking at the Hansard[5] of the House of Commons (*Educational Settings - Hansard - UK Parliament*, 2020), it showed no record of debate around the matter, nor about the announcement of the use of a process to calculate grades. While such debates did appear in wake of the contestation led by students after the filing of grades, the absence of such debates could be due to the many difficulties faced by the government in 2020 due to the covid-19 crisis, which might have set aside such discussion within the House of Commons.

Already, following on the relevant social groups identified earlier, we can extend on the SCOT method (**Bijker,** 1995) to identify the producers and the ordinary users. In our case, OfQual are the producers while the boards are the users. That underlines the small role played by the students within the construction of the technology and the negotiations during its development. Moreover, we can already observe that the algorithm was built upon the

[5] Those are the transcription of debates going on in the House of Commons in England.

objectives of its makers. This goes in line with the argument made by **Cardon** (2019) that algorithms are embedded with the interest of their makers, but also on the system from which they stem.

Building on the education system:

In a report from the 3rd of April 2020 (OfQual, 2020a), OfQual announced that they were going to ask some resources from centres: an assessed grade for each student for each subject, a rank order of students within a subject with the highest attaining student on top, and a declaration from the head of each centre. They asked for these resources before knowing what they were going to do with them as they wanted to act in a timely manner. Those were added to the resources that were already available and used by OfQual each year.

Indeed, as predictions about the general average of students are made each year to decide on the grade distribution[6] (**Kelly**, 2021), OfQual already had access to a lot of historical data about the history of each centre. They knew the final grades for each student's A-Level as well as the prior attainments of most of them. According to OfQual, the most "readily available form of evidence regarding students recent performance [is] teacher estimate" (OfQual, 2020c, p.13).

But OfQual asked for the ranking of students as they deemed that the usual predictions were not precise enough (OfQual, 2020c). To affirm that, they used previous studies on the matter 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 relative to each other's (*relative accuracy*).

The studies they referred to were made by the Cambridge universities about the prediction of teachers (OfQual, 2020c). They were mainly about teachers predicting grades for students in general and not so much about examinations specifically. They also used tests on their historic data in order to support that *absolute accuracy* was less optimal for their objectives. According to them, teachers were imprecise in about half of the cases. And most of those times, the teachers were too optimistic.

[6] In a year with examination, predictions are used to decide to what grade correspond each mark within each exam board and within each subject.

Using those sources as a reference, OfQual argued that using grades predicted by teachers through absolute accuracy would thus 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, one reason for that inflation could for example be that teachers anticipated the standardisation process and raised their students' grades above their expectations.

Also, using CAGs directly would have led to inequalities between centres according to OfQual (OfQual, 2020c). They argued that, as they were not able to properly form teachers to produce accurate standards, there would (1) be too many disparities between centres and (2) variation between years would be too high and lessen the values of higher grades for the 2020 cohort. As the highest of CAGs and calculated grades were adopted in the end due to the U-turn, it showed that (1) the disparities between centres did not widen but (2) the general average did exceed national year-on-year inflation (OfQual, 2021).

In order to follow the directives set by the Department for Education, OfQual was thus tasked with putting in place a process to smooth the CAGs so that they would be in line with historical distribution of grades. In order to do so, OfQual came up with the following set of objectives for their grading system:

> ". to provide students with the grades that they would most likely have achieved had they been able to complete their assessments in summer 2020
>
> . to apply a common standardisation approach, within and across subjects, for as many students as possible
>
> . 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
>
> . to be deliverable by exam boards in a consistent and timely way that they can quality assure and can be overseen effectively by Ofqual
>
> . to use a method that is transparent and easy to explain, wherever possible, to encourage engagement and build confidence"
>
> (OfQual, 2020c)

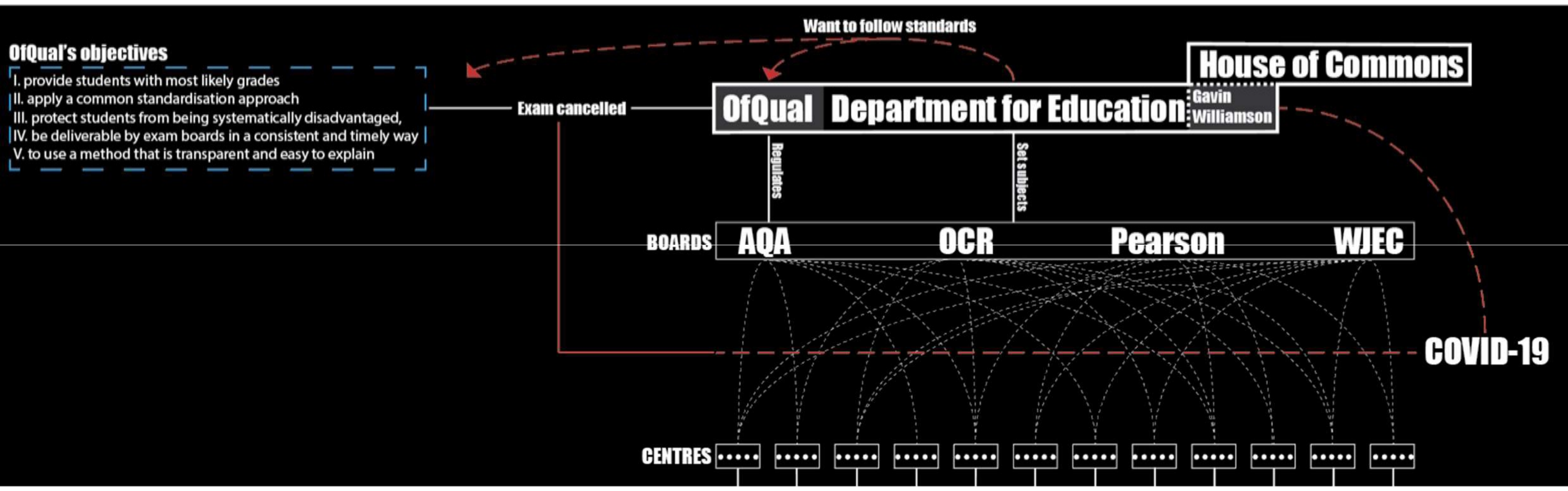
OfQual organised public consultations online around those objectives and their general approach (OfQual, 2020b). And a majority of the feedback was positive. As we will see throughout our analysis, those objectives led to some clear directions to the project. But we can already see that the system was built with the clear objective of reproducing existing patterns by promoting an approach that follows general grading distributions. That left OfQual in a position that did not allow for a lot of freedom to overcome existing structural discriminations.

In my interview with one of the directors of the project, I asked if he believed the result could have been profoundly different if a more complex system had been used with the same limitation. According to him, as they were trying to reproduce the historical distribution of grades, it was very difficult to overcome general differences between centres (**interview**). This actually led to a system that reproduced the general disparities between centres that are imbued with differences in socio-economic factors (OfQual, 2021).

Something should be noted here about the relation between OfQual and the department for education. Many times, within the documents produced by OfQual, they underline that most of their choices come from the willingness to follow the rules from the government (OfQual, 2020c). During the interview with one of OfQual's directors, he often referred to the implementation of a policy (**interview**), as if they were taking political decisions in opposition to what they were doing. In a way, he suggested that OfQual only enacted the decision from the government in an apolitical way.

Of course, as we saw in STS literature, we should not trust that a system could be non-political. In this case, we see that the many decisions that OfQual already took to this point had a major impact on the outcome of the system and the lives of students. Following on the argument made by **Seaver** (2017) that algorithms are entangled within their societal environment, we can identify here a point of entanglement as the objectives are rooting the algorithm within its environment.

Even if these objectives are strongly dependent on the incentives from the government, choosing these and prioritising them as such is a political choice made on the basis of some *a priori* on society. Even if those decisions are sometimes built on the premise of scientific literature, choosing to follow those objectives over others is still a subjective choice. This is a first step in the imbrication of the algorithm within its societal context.


Want to follow standards
OfQual's objectives
I. provide students with most likely grades
II. apply a common standardisation approach
III. protect students from being systematically disadvantaged,
IV. be deliverable by exam boards in a consistent and timely way
V. to use a method that is transparent and easy to explain
Exam cancelled
House of Commons
OfQual
Department for Education
Gavin Williamson
Regulates
Set subjects
BOARDS
AQA
OCR
Pearson
WJEC
COVID-19
CENTRES

We might illustrate the impact of the covid-19 as such. This shows that the cancellation of examinations by the House of Commons in response to the difficulties brought by the virus removed the relation between centres and exam boards. Yet, the links between the boards and the government were maintained as no decisions were taken that had an impact on their relation.

We can see that the *translation* from an objective of comparison to examinations is broken by the politicians' decisions. As boards cannot act as an OPP anymore, the system made by OfQual is made in order to take upon that role. But as it happens between OfQual and the boards, teachers and students are scrapped from the negotiation. The process of *interessment* put forth by OfQual to ensure the comparability of the system in regard to the objectives of the exam boards to deliver certification thus fail to take into consideration students as well as teachers.

This also means that the general objective of the system as a whole is changed as OfQual is tasked with the role of assessment. Its original objective of ensuring comparability is changed into the development of a grading system. As we will see throughout the confection of the model, the initial objective of OfQual is going to nudge the whole system in a certain direction.

We will later go in depth within the technical choices made by OfQual to see how they were negotiated. But first, we will focus on the idea of building transparency into the system. This was one of the objectives but, as we saw earlier with **Christin** (2020) it is difficult to build algorithms that are not complete black boxes, and the black boxing of the algorithm plays an important role in allowing or not negotiations around and with it.

Making the algorithm blurry

As part of their objectives, OfQual tried to be transparent to build confidence around their system (OfQual, 2020c). To do so, they developed some strategies to give access to their model. A lot of reports were made available online in which they break down their approach in order to make it accessible for people concerned but also for people who might want to build research based on their experience. The interim report from which most of this analysis stems is part of that idea.

Still, those reports were not always easy to find, buried within all the information produced by the government at the time. For example, the decision to put in place a standardisation approach was made public through the Hansards of the House of Commons, meaning that finding such information requires the knowledge required to browse amongst all the web pages of the government. Moreover, some of the subjects covered within those reports were sometimes very technical and took patience and determination to decipher.

Also, the report explaining the inner workings of the algorithm, its consequences and the step-by-step decisions made by OfQual was published on the 13th of August, on the same day as the results. That had the consequence that students, had they gotten the patience and determination to do so, could not fully grasp the workings of the algorithm in order to build confidence in the system.

Yet, OfQual tried to make their system available and understandable in other ways. They made a video explaining how the calculation was going to work (**Ofqual**, 2020d). That video was a synthesis, available on YouTube, of the working of their algorithm so that students and parents could understand what was going to happen. Yet, the small number of views (55,207) compared to the number of students being graded (around 250,000) suggest that it was not very efficient.

Then, following on their objective of transparency, they made the code available for anyone to see it (*Summer 2020 Code Used to Grade Qualifications*, 2020). That code was not really the one used in the calculation process as it was stripped from its datasets and its environment as a whole. It was more of a skeleton of the model that was given to exam boards to implement. Yet, OfQual wanted to make that code available in order to allow other people to study and work on it.

As we have already seen, public consultation was also organised around the objectives they should pursue. Those happened between the 22nd of May and the 8th of June. OfQual asked for feedback on different questions (OfQual, 2020b). Out of 12,623 responses they received, 1,939 were from students. Here again, the small number of students who responded could be an indication that their transparency policy did not lead to a bigger engagement from the people concerned. But it might also be possible that people were in general agreement with those objectives but did not feel the need to express their support.

Those methods to render their process as transparent as possible were thus not as efficient as they wanted them to be, following their objectives. Using the characteristics of black box algorithms identified by **Christin** (2020)[7] we can identify how the transparency promoted by OfQual failed to address the opacity of their system.

First, while OfQual did render their objectives available, the technical process, as well as the code, were kept in *secrecy* until the results were issued. Then, as we will see next, the choice of a computational approach built *technical illiteracy* within the system as not all students had the skills to study the code. Finally, the *size* of the system was also substantial as about 700,000 A-Levels had to be delivered, which brought more difficulties to understand the system. Only the *untelligibility* was not really present in the system as it was always possible to redo the calculations made by the algorithm.

But as OfQual failed to build transparency into their system, we see that the process of *interessment* of students falls apart. As students were not part of the relevant social groups in the implementation of the algorithm, they have created a blurry system for them where they could access some information but, at the same time, the algorithm is still a black box for them. Moreover, by failing to enrol students in the assemblage of the algorithm, OfQual failed to convince them of the necessity of the implementation of their model compared to the usual system.

We may also wonder if a complete transparency would have fitted their goals as they were trying to provide a simple and rapid solution (OfQual, 2020c). Indeed, they may not have taken the time and resources necessary to teach each individual the inner workings of their algorithm. Also, as we will also see at the end of this analysis, OfQual did not view their process

[7] *Intentional secrecy, technical illiteracy, untelligibility* and *size.*

as a black box. They may thus have overlooked the impact of their approaches in regard to students by thinking that they were simply emulating the usual actions of exam boards.

That mis-conduct of transparency also raises the question of implementing such a process within a democracy. As the process of assessing pupils is made hidden from all in the form of a black box algorithm, the power to grade students shift from accountable members of society to a heterogeneous artefact that only some can comprehend and tinker with. And as will now see, tinkering with the model can lead to dire consequences.

Moreover, and as we will see throughout the upcoming parts of this work, the blurry opacity of the system meant that it was harder for students and teachers to engage with it. In this next part will see how the choices made by OfQual, while considering a model to replace examinations, lead to more opacity from the algorithm and the further muting of students.

Finding an alternative

As it was asked of OfQual that grades followed a similar profile than previous years (**Williamson**, 2020), they studied the possibilities for different types of standardisations. They chose to use a meso-level[8] standardisation model as they were asked to standardise at centre level. That meant that they gave more weight to centres than to the system as a whole or students individually. According to them, that approach was the most relevant as it allowed them to be closer to the national average without having to produce a model to be applied for each student individually (OfQual, 2020c, p.34). But before developing their approach, they had to overcome some issues that could lessen the precision of their system.

Obstacle:

First, each year, some students choose to or are not able to pass their exams. As it would be too difficult to implement those absentees in their models, and in a spirit to give to students the benefit of the doubt (OfQual, 2020c), OfQual chose to ignore the issue, thus giving grades to students who would not have passed their exams. During the interview with one of OfQual directors, he confirmed that it was an opportunity to give students the benefit of the doubt while also avoiding bringing more complexity within the system (**interview**, R26). But a drawback

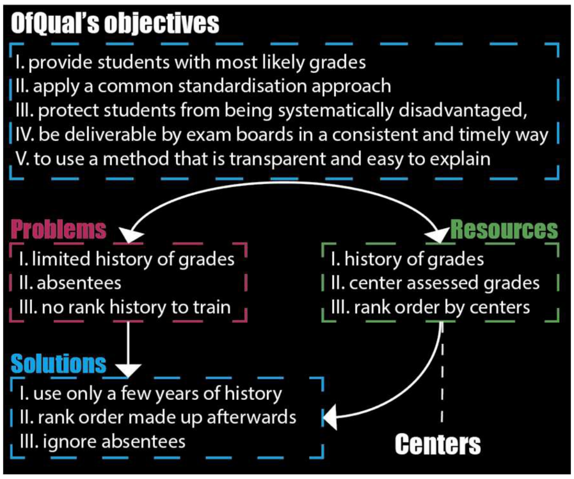
[8] As opposed to micro or macro level standardisation, levelling student individually, or as a whole respectively.

of that was that the historical data used to train with the model was less appropriate as it had absentees for other years.

Then, the data available historically was limited due to reforms on education in 2016. They re-worked the subjects of A-Levels meaning that some of them were cancelled while others were created. Due to that, they only had a few years of data available in some subjects to train their models. They chose to use only three years of data for training in each subject in order to avoid having differences in precision between them. This meant that the model lost in precision as it was trained on less data than it could have.

Finally, there was no available ranking for previous years as it was not asked to teachers in previous years. In order to still train their models with the ranking system, they built ranking *a priori* based on the A level achievements of students. Students with the highest A level being put at the top of the ranking. Of course, these ranks were not representative of what teachers would have done as it was based on the grade that the student actually got and not ones predicted beforehand like they were in 2020.

Taking those limitations in consideration, OfQual then started to develop different approaches. Together with statistical and educational experts from different governmental institutions and exams boards, OfQual developed eleven models to standardise grades. During my interview with a director from OfQual (**interview**, R30), I also learned that an advisory board of statisticians apart from the government was also put in place in order to answer questions regarding the development of their model.


OfQual's objectives
I. provide students with most likely grades
II. apply a common standardisation approach
III. protect students from being systematically disadvantaged,
IV. be deliverable by exam boards in a consistent and timely way
V. to use a method that is transparent and easy to explain
Problems
I. limited history of grades
II. absentees
III. no rank history to train
Resources
I. history of grades
II. center assessed grades
III. rank order by centers
Solutions
I. use only a few years of history
II. rank order made up afterwards
III. ignore absentees
Centers

Representation of OfQual limitations and solution to achieve their objectives.

Once again, this shows how the algorithm was rooted within its societal context with precise objectives. As it was argued by **Kitchin** (2017), unpacking the algorithm allows us to see just how it is imbued within social relations. Moreover, the choice of using a computational method induced a form of *technical illiteracy* (**Christin**, 2020) which meant that students were less likely to be formed to understand the inner workings of the method. But as those approaches were still intelligible, OfQual was able to test each of them.

Auditing the algorithms:

The eleven models were based on three main approaches. We will later go in depth into the chosen model, but for now we will quickly see the distinctions between each of them. Some of the models were lapping between two types of approaches. But nonetheless, OfQual divided their models into three main categories.

The first set of models were based on what OfQual calls a *direct centre approach* (DCP). Those approaches directly used the history of each centre to identify a model. Those approaches then compared the model of each centre with the CAGs and 'corrected' them if they were either too high or too low.

The second set of models were based on marks (MBR). Those models were meant to build more closely on the history of each candidate. A prediction was made for each student individually based on its history. All those predictions were then put in a decreasing order that was applied to the ranking order provided by centres. The first student in the ranking would thus obtain the highest predicted mark for a student of its centre.

The final set of models used grades (GBR) as its main source of data. These models identify the probability for each grade to be represented within each centre based on their history. It then applies the rank in order to the proportion identified. For example, if the model predict that they will be five students attaining the grade A*, the first five student of the ranking will get the grade A*.

More approaches were considered but, due to the limitations we saw above, those were too hard to put in place, mainly due to the lack of historical data (OfQual, 2020c). As argued by **Cardon** (2019), the ways through which algorithms are imbued with their creators' values is in part due to their technical choices. Here, we can identify how the choice of those models in particular showed a leniency towards the reproduction of the existing grade distribution as those heavily rely on past grades to make their predictions.

Then, in order to choose what model to use, OfQual development team tested their accuracy through statistical analysis, which is a form of algorithmic auditing as we will see. To do so they tested each model with the data from 2016 to 2018 to study their accuracy to predict the grades of 2019. Where it was needed, they used the 2019 ranking built *a posteriori* on the basis of students' grades. Those tests were realised amongst all subjects for both A-Levels and GCSEs.

As a result of those tests, it appeared that the DCP approaches were all amongst the most accurate. One of the marks based one and one of the grades based one were also attaining similar levels of precisions. More approaches were in the same range of predictive accuracy for A-Levels, but they were lower for the GCSEs. As OfQual needed to develop a unified approach for all exams, they ruled out those approaches to only keep 5 of them.

To decide between those five approaches, OfQual studied their practical aspects. As the model was to be applied within each board individually, the chosen model should be one that can easily be replicated and communicated. In that regard, OfQual argued that the DCP models 1 and 3 were the least complex to put in place (OfQual, 2020c, p.60). According to OfQual, those methods were less demanding in computational resources and were easier to replicate within the existing systems of exam boards. This was due to the fact that the MBR and the GBR methods, as well as the DCP second method[9] were more relying on following mathematical procedures than fitting models.

Moreover, in order to compare the two methods, OfQual studied their impact on equality within each centre. They wanted to identify the impact of their method on two main aspects (OfQual, 2020c, p.62). First, how the demographic of a centre impacted the accuracy of predicted grades. Then, if changes in the demographic of a centre impacted the accuracy of predicted grades. Finally, they wanted to know which of the two DCP models was the least biased.

To do so, they created seven socio-economic categories based on the protected socio-demographic in England[10]. They also created two-ways categories by cross-analysing ethnicity with both gender and free meals eligibility. They considered cross analysing other categories but that would have led to cohorts that were too small to provide relevant statistical information.

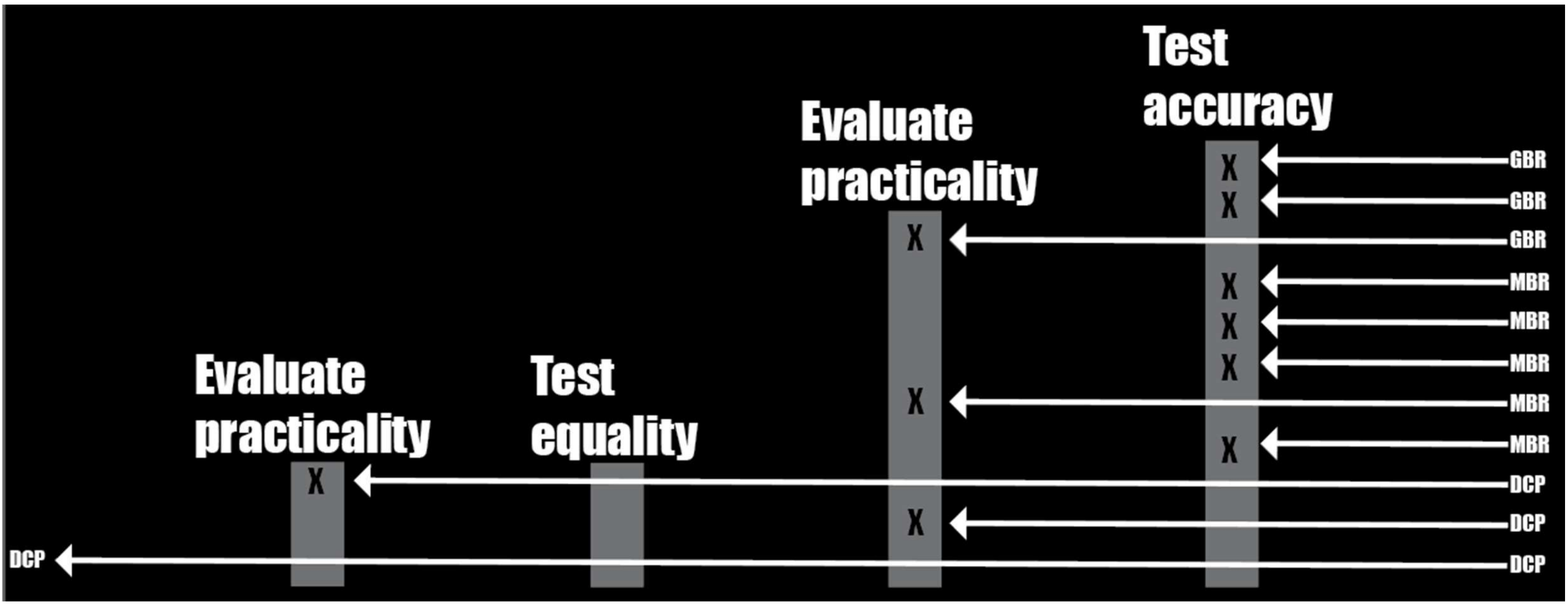
They then proceeded to study how those categories had an impact on the prediction of the following three A-Level subjects: biology, French and religious studies. In their reports, OfQual does not go in depth in the choice of those subjects and its consequences. They simply explain that those three subjects represent different size of cohort and various subjects in order to provide a variation to their analysis (OfQual, 2020c, p.53). Also, while OfQual studied the

[9] That method was a fusion between MBR and DCP, meaning that a MBR approach had to be performed before applying the DCP method.

[10] Those were as follow: proportion of female students; proportion of student with special educational needs; proportion of students with English as additional language; proportion of student eligible for free schools' meals; proportion of students in each tertile of the IDACI score; proportion of non-white students; a breakdown of non-white student in more precise categorization.

statistical impact their model could have within centres, they did not consider the general inequalities between centres.

Yet, whilst it is unclear if other subjects might have led to other results, OfQual concluded that their methods had no impact on the different categories. They concluded that the impact of the socio-demographic composition of centres had no impact on the accuracy of the prediction and the year-on-year variability (OfQual, 2020c, p.75). As they were not able to argue in favour of one of the models on the basis of equality, they argued in favour of the first DCP approach in regard to the practicality of implementing the approach. From now on, we will refer to this approach, which was selected by OfQual to be the one used for the grade calculation, simply as the DCP approach for clarity purposes.



Summary of the decision process behind choosing the model.

In regard to what we have just seen, we can argue that choosing this method was not made on the basis of its accuracy alone. Whilst some approaches were discarded because they were not precise enough, the choice of the approach to use was decided on based on the practicalities needed to implement it. In regard to the study of biases between the approaches, it did not result in the elimination of any approach and was deemed as not relevant to the development of the model in the end.

This shows that what was essentially an *audit* of their algorithm did not lead to a full comprehension of the discrimination at play. Or at least those *audits* are strongly dependent on the making of categories beforehand. This is due to the limitations identified by Christin (2020) of such an approach. Indeed, it shows that technical analysis, without regard for the system as a whole, is not enough to engage critically with the algorithm.

As we saw before, this was supported in a recent paper from the algorithmic justice league, leading research group on algorithm auditing (**Costanza-Chock et al.**, 2022). In their paper, they argue that audits should pay more attention to the structural forces at play around the algorithm. In line with **Christin** (2020), they argue that quantitative analyses of algorithms are not sufficient if the general system is not considered.

Moreover, the general process of choosing the algorithm shows how the technical choices were made to fit an existing model. Following the analysis of politics within artefacts (**Winner**, 1986), we could thus argue that the model was *"*stacked in advance to favour certain social interests and that some people were bound to receive a better hand than others" (p.26). In our case the social interest was to keep in line with standards and, as we are going to see, some students were bound to receive a better hand.

The emulation of the examination model

The algorithm was thus built following the DCP model in order to simulate the exams that had been cancelled. We are now going to study its particularities as well as the consequence of the approach on the general calculation of grades. We will then see how it was implemented within each exam board. Finally, we will study the consequences on the grades of the whole process.

Reproducing the historical model:

The way the model works is by first measuring the historical distribution of grades within a centre for a subject to produce a *prediction matrix*. On the basis of the matrix, the model analyses the distribution of previous attainments in past years as a whole, from 2017 to 2019. It then calculates the distribution of prior attainments within the 2020 cohort. Based on how many students were actually matched with their prior attainments, the model adjusts the initial distribution of grades to be more in line with previous achievements and the history of each centre.

Finally, the students are matched to the prediction on the basis of the ranking. If there are 12% students predicted to get a grade A*, the first 12% of the ranking get the grade A*. In order to avoid having students between two grades[11], marks are calculated on the basis of the predicted distribution and the history of each centre. The student's grade is then settled on the basis of that predicted mark.

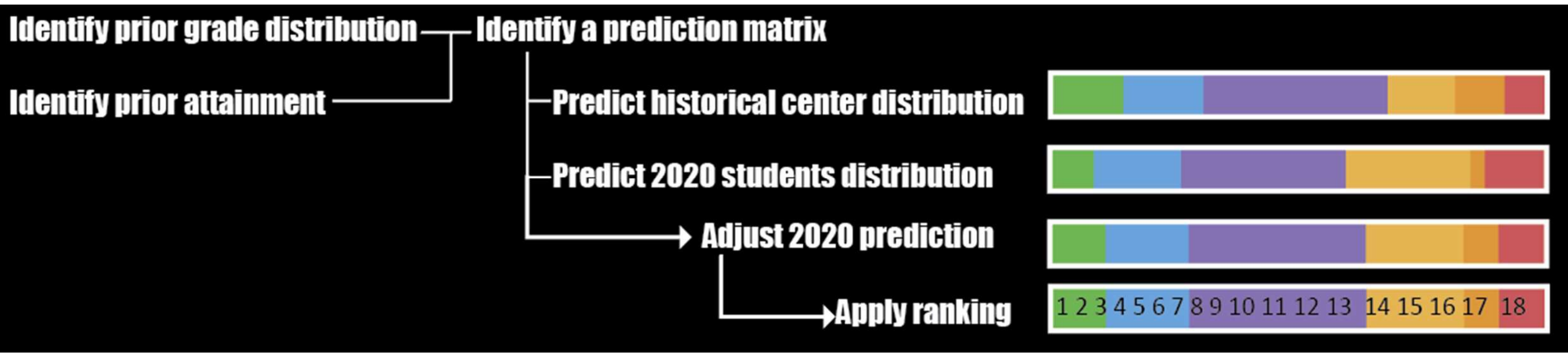

Identify prior grade distribution
Identify a prediction matrix
Identify prior attainment
Predict historical center distribution
Predict 2020 students distribution
Adjust 2020 prediction
Apply ranking
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Illustration of the DCP model

It is through that process that the statistical model complexifies. As we saw earlier, we define algorithms as a complexified form of code **Panch** (2019). Here, the process of prediction carried by the code is characteristic of that change from code to algorithm. It is also the moment when the algorithm gains most of the characteristics of a black box as formulated by **Christin** (2020). As a consequence, the system put in place by OfQual becomes opaque and what was presented as an objective system becomes untouchable.

In regard to the categories identified by **Christin** (2020), we already underlined how the system was imbued with a form of *secrecy* and the *technical illiteracy* brought by the general approach. But with the complexity brought here, the *technical illiteracy* increases and the *size* of the system, applying the model to each centre for each subject, renders the system more opaque. The last characteristic being the *untelligibility* of the system, to laypersons as well as to its maker. But as we will see, while the system was not intelligible to students, its developers still understood it.

In regard to the DCP model, its implementation meant that the students' calculated grades strongly relied on the proportion of students that achieved each grade from 2017 to 2019 within each centre. The whole process gave more weight to the history of each centre than to the CAGs and ranks produced by teachers within each centre. That is a direct consequence of the government's injunction that we saw earlier. Indeed, this was made in order to ensure that the general average of each centre, and thus the national average, would not change too much.

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

The direct consequence of that approach was that it reproduced existing patterns. In their interim report (OfQual, 2020c), OfQual studied the potential inequalities imbued by their system. They demonstrated that the points obtained by each protected category identified while testing for inequalities were in line with previous years variations. But that meant that attainment gaps between some populations of students were maintained.

For example, by looking at the tables given by OfQual (OfQual, 2020c), we can see that student 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 student who do 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. Whilst it showed that OfQual was right in stating that there is no evidence that their system introduced biases (OfQual, 2020c, p. 179), it also shows that they failed to consider the system as a whole.

Still, it is important to note that the calculation process slightly reduced disparities on the basis of socio-economic status for grade A and above compared to the note given by the CAGs (OfQual, 2020e). Yet, if OfQual had applied CAGs directly, the disparities on the same basis would have stayed the same overall. This shows that the standardisation process removed less higher grades for students with a low socio-economic status. That tendency reverses from grade C and under. That corresponds with what OfQual expected from the prediction by teachers, which motivated the use of a standardisation process.

Overall, the algorithm successfully reproduced existing disparities between the protected categories that were identified earlier. Nonetheless, analysis of the grade given in 2020 on the basis of the CAGs, after the U-turn[12], there is no evidence that the CAGs were more or less biased than the usual disparities over the years (OfQual, 2021). The main difference brought by the CAGs being the overall higher distribution of A* and A grades. That shows us that while teachers might tend to over-estimate their students in general in regard to their exams, they sadly are very good at reproducing discriminatory patterns.

[12] Those grades were the highest for each student between CAGs and calculated grades.

Filling the gaps:

But while the algorithm was applied to all A-Levels, two categories of students were not fitting into OfQual's model (OfQual, 2020c). Indeed, as it relied strongly on the history of centres, it meant that students isolated from the system or within too small centres could not be assigned with accurate results. Those were the private candidates and the small cohort.

Regarding the first, they refer to a centre only for the examinations but not for the formation. In order to avoid those students from messing up the rank order of the students who actually 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.

The other exception was made for small entries. As small cohorts 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 CAGs directly as using their model would not be representative of the prior attainments from students. OfQual measured that in order to keep in line with the general distribution of grades over the years, centres under 15 could be handed their CAGs directly. All centres with entries over 15 students by subjects had to go through the standardisation process.

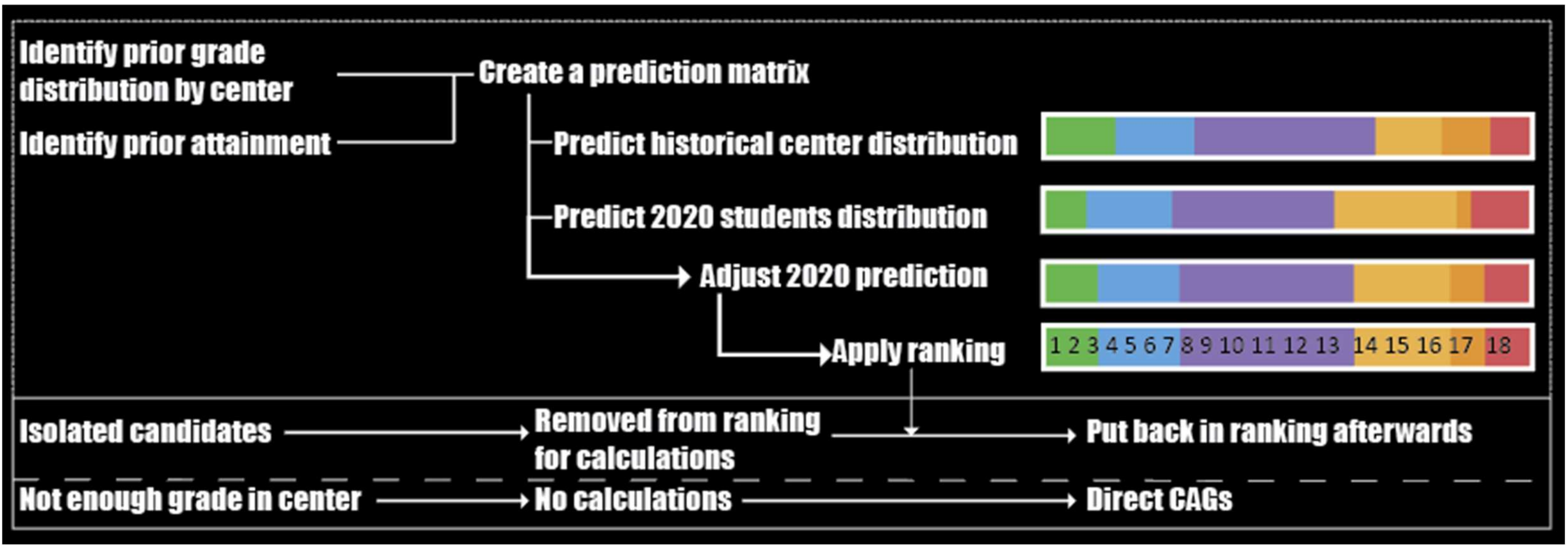

Identify prior grade distribution by center
Identify prior attainment
Create a prediction matrix
Predict historical center distribution
Predict 2020 students distribution
Adjust 2020 prediction
Apply ranking
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Isolated candidates
Removed from ranking for calculations
Put back in ranking afterwards
Not enough grade in center
No calculations
Direct CAGs

Illustration of the model as a whole.

According to **Kelly** (2021) this was an unfair advantage given to wealthier centres (as in not funded by the state, not to be mingled with private students). Indeed, he argued that small centres were mainly richer, fee-paying centres, which already selected students on the basis of their prior attainments. This goes in line with the argument made by **Winner** (1986) as the algorithm is designed in a way that reproduces patterns of oppression.

This whole process could be viewed in terms of algorithm refraction according to the meaning of **Christin** (2020). Indeed, she invites researchers to look at how algorithms are applying a form of refraction to society to produce a deformed picture of their environment. In this case, OfQual refracted the general capabilities of students into a distribution of grades according to general standards. But we can also see that it mainly refracted some aspects of society, with some students being unaffected by the model.

Unpacking the full assemblage of the algorithm as encouraged by **Seaver** (2017) thus showed how the algorithm had an impact on the population of students as a whole. But it also allows us to see how the process as a whole is entangled with the historical distribution of grades. Which helps us to account for how the algorithm was shaped by its environment and how it reproduced existing biases of the educational system.

The reproduction of historical biases by algorithms is actually a risk that was identified by the centre for data, ethics, and innovation (CDEI) of the English government, whose chair was also the chair of OfQual at the time. But as they put it, the CDEI "has not had any direct role in assessing OfQual's approach" (*Review into Bias in Algorithmic Decision-Making*, 2021).

Yet, as OfQual had successfully designed an algorithm in order to reproduce previous years grades, they still had to implement it within the exam boards. This was necessary as OfQual was not meant to take the role of examiners as they could not deliver the certifications. Exam boards thus had to implement the model in order to be able to give the results by themselves.

Reproducing the recipe:

Once OfQual had developed its approach, they had to put it in the form of a code that would be applicable by the four exam boards individually. To do so, they chose the coding language 'R'. According to the director I interviewed, they chose that language as it is pretty standard and already used within OfQual for their year-on-year analysis (**interview**, R42). He also told me that they could have used other languages such as SAS[13] but it was not necessary given the lesser complexity of their system.

Once their code was developed, they sent it to each of the exam boards for them to implement in their own system. Not all boards were using the R language, and some thus had to tinker a bit with the original code. According to the director interviewed, OfQual spent a lot of time ensuring that the way their model was implemented within each board was actually equivalent. They did so by ensuring that they obtained the same results over small samples (**interview**, R45).

The calculation was made on the whole national level. While it usually depends on exam boards to put in place their own assessment methods, they had to harmonise under the standardisation process in 2020. However, some centres change exam boards between the years to mark the same subject. As we saw earlier, that could be due to the services offered by the board or the preferences of the teachers.

Implementing OfQual's code within only their own history would thus have been very difficult for the exam boards. In order to avoid having to implement such complexity in their system, OfQual built and made available a database with the national history of grades for all centres. That allowed boards to put in place the model themselves on the basis of that dataset. As I learned in the interview, it is actually common for OfQual to put in place such a dataset for their year-on-year analysis.

At this point, boards took the CAGs and ranks from the centres and applied the model where it was applicable. They then uploaded the calculated grades like they usually do, through a common platform for all boards. And on the 13th of August, students received their grades. In parallel, OfQual released their interim report in order to explain in detail the standardisation process. They also published a review of the grades awarded.

[13] Statistical Analysis System.

That step is crucial in terms of *arrangement* and *interessment* in algorithms. As **Christin** (2020) argued, algorithms could be studied in terms of how the negotiations around them happen. In our case, we have identified how the negotiations between OfQual and the boards led to the implementation of the algorithm, but also to the muting of students as a whole.

At the beginning of our analysis, we identified that the boards were acting as obligatory passage points (OPP) in the English educational system. We then saw how that OPP switched to the algorithm. By implementing the algorithm within each centre, we can argue that the boards are the OPP again. But while that could suggest that the network is stabilised as the OPP was back to where it started, we are going to see that the change of nature of the OPP, with the implementation of calculated grades, led to new negotiations.

Wobbly return to normal:

As a consequence of the algorithm, 41.3% of the CAGs had been tampered with to fit with the general distribution of their centres. 39.1% of the total being downgraded. As we demonstrated earlier, the results of the calculated grades showed no anomalies in terms of equality between students in regard to previous years. Its main consequence was to lower the grades that were given by the teachers. The algorithm had thus fulfilled its goal of providing students with grades in regard to standards.

But on the 13th of august, as students received their grades the same way they would usually do, contestation appeared. Quickly, *#fuckthealgorithm* was used as a rallying cry for students on twitter (**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 ≠ stupid” (**Benjamin**, 2022).

But, while it was true that the algorithm indeed reproduced existing discrimination, it is not on that basis that student protested. Indeed, as **Kelly** (2021) underlines, there was a belief that students from lower-socio economic backgrounds were at higher risk of seeing their grades downgraded. As we saw, that was the case for lower grades, but it was not true for the calculated grades as a whole.

According to **Benjamin** (2022), the protests were not really about the downgrading of students from lower socio-economic backgrounds. He argued that the protests were as much about the usage of an automatic process as it was about the reproduction of existing social injustice by the government. According to him, the protests were against the algorithm, but they were in fact targeting the government that put them in place.

As a matter of fact, by putting in place the algorithm as an OPP 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 as the system was simply reproducing itself without their implication. The shift of OPP had thus opened new negotiations around the stabilisation of the network.

A precise sociology of those protests might reveal exactly what pushed the students to go into the street and to chant in front of the Department for Education. Nonetheless, due to those protests and the pressure exerted on the government by the public opinion, they decided to go back on their decision and give students the highest grade between their CAGs or the calculated one[14].

As students were muted during the development of the algorithm, that led to new negotiations when those entered in relation with it. The government was thus offered with two main choices. Firstly, they could have enforced the role of the algorithm as a tool to replace the examinations and teachers assessment. Secondly, they could accept the expertise of teachers and apply their CAGs directly. But as they chose the second option, that meant that the process as a whole was rendered useless.

Moreover, on the 19th of August the university application service (UCAS) announced that students who saw their grades downgraded by the algorithm could now re-apply with their new grades. The logistics of such manoeuvre was left to universities. But having already gone through the whole process once, most of the curriculum was already booked. The government thus had to remove the cap imposed on student numbers in universities to allow newcomers. Finally, on the 15th of September, Gavin Williamson was removed from his role as education secretary as prime minister Boris Johnson reshuffled his team after the first round of the pandemic ('Reshuffle', 2021).

[14] Students with a higher calculated grade than a CAG represented 2.2% of the system.

Fitting the system rather than fitting students

Throughout our analysis, we saw that OfQual official reports chose to refer to the system as "the model", "the approach" or "the standardisation" rather than calling it an algorithm. During the interview with the OfQual director, I learned that they refused to refer to their system as such as they felt it was an unfair representation of what they did. According to said director, their system was far less complex than an algorithm. It would then not be appropriate to refer to it as such as it had a small degree of autonomy compared to algorithms (**interview**, R21).

But as we saw, the automation of the process along with the opaqueness of the system are two key points of algorithms. Indeed, we saw that the general complexification brought by the computational method to apply the model to every centre corresponded to the transition from code to algorithms according to **Panch** (2019).

Moreover, we found that the system was characterised by its opaqueness, which is another key aspect of algorithms according to **Christin** (2020). We demonstrated how the computational approach led to *technical illiteracy* from the students. How the lack of transparency in the development is a form of *intentional secrecy* and how the *size* of the model rendered it difficult to understand it overall. Only the *untelligibility* cannot be used to qualify the system as OfQual claimed it was always possible to redo the calculations made by the algorithm.

Identifying algorithm as such is important as the director also said that one other reason they refused to refer to their system as an algorithm was to avoid the regulations from the ICO (**interview**, R21). The ICO is the organism in charge of regulating information rights in England. It also issued strict rules on the implementation of automatic processes. They thus avoided qualifying their system as automatic as that would have led to a review by that organisation.

As the results were announced and the students were protesting, the ICO was made aware of the case and stated the following: "Ofqual has stated that automated decision making does not take place when the standardisation model is applied, and that teachers and exam board officers are involved in decisions on calculated grades." (*Statement in Response to Exam Results*, 2020).

That interpretation by OfQual, and thus the ICO, of the algorithm as being non-automatic because it is surrounded by human intervention could be due to its *intelligibility* as they were able to comprehend each step taken by the algorithm, they thus viewed the algorithm more as a replacement of human actions rather than an automatic decision process. But of course, we saw during our analysis that there is actually an automatic decision process taking place in order to calculate the predicted grades of each student. While it is true that that process is surrounded by human involvement, we saw that it is not sufficient to say that the decision in not automatic, and that it would be misleading to say it is not an algorithm.

At the end of their statement, the ICO said that they were going to monitor the case, but as the algorithm was scrapped, it does not seem that they pursued. But whether OfQual did it intentionally or not, what is of interest to us is to see how the "model" of OfQual actually became an algorithm.

The appellation of the algorithm appeared after the grades announcement as students were contesting the results (**Benjamin**, 2022). It was also used by journalists and politicians to refer to OfQual's code afterwards (**Hao**, 2020). It was also qualified as such by OfQual themselves later (OfQual, 2021). It shows that there is a clear difference between the statistical models translated into code and what we call algorithms after the announcement of grades.

That moment, when the code is enacted by social practices to become an algorithm is the reason why **Seaver** (2017) encourages us to view algorithms *as* culture instead of *in* culture. Here, the algorithm becomes qualified as such from the moment it is enacted by the different actants. The algorithm is thus not a piece aside from a system that was built from scratch but rather one of the pieces of that system. As we identified, the whole process of developing the algorithm was strongly dependent on the social system around it, and thus strongly related to it.

And in the same way as it was not built from scratch, its entanglement within the system means that it did not instantly disappear from the system as a consequence of the U-turn. The OfQual director interviewed explained that the system used in the summer of 2021, as exams could still not happen, was built on a similar approach. But their new model was built more closely to the CAGs and gave results similar to those awarded after the U-turn, with the objectives of deflating grades over time to go back to the standards prior to 2020.

Another interesting aspect of the algorithm is that, while it succeeded to fulfil its goal of reproducing the standards, it gave visibility to some existing disparities from the educational system. By being so close to the existing biases, it rendered visible the disparities between centres. Due to the algorithm, people started contesting those systemic discrimination.

During the interview, the director jokingly argued that "blaming social discriminations on exams is the same blaming climate change on the thermometer" (**interview**, R15). It might therefore be that, as thermometers provide useful data to analyse climate change, the algorithms provided with evidence of discrimination within the system. This is in line with the analysis made by **Benjamin** (2022) as he argues that protests against the algorithm were actually more directed to decision makers and the educational system in general.

Mapping the OfQual algorithm

Throughout this work, we tried to render visible the different steps of the making of the OfQual algorithm. The map at the end of this section is a summary of those rendition to offer a view of the development of the algorithm. It is a synthesis of the different steps we identified throughout our analysis that allows us to have a general view of the algorithm and its consequences.

First, at the bottom right, we have the students who are being graded. Those are in cohorts of various numbers. They are assembled within centres and assessed by teachers. As we saw, each centre refers to an exam board for the examination of A-Levels. But in response to the covid-19 pandemic, those examinations were cancelled by the government. OfQual thus had to come up with a way to provide a way for the exam board to provide grades to students in order to fit with the general system of education.

To do so, they came up with a set of objectives that they wanted to follow. But in regard to the general incentives of maintaining standards and comparability, and facing some issues, those objectives were prioritized. Indeed, we saw that the application of a common standardisation approach was in fact the main goal of the algorithm. On the contrary, avoiding students from being systematically disadvantaged was in fact impossible as the algorithm was made to reproduce the current grade distribution, along with its systematic discriminations.

From there on, the algorithm was imbued with those goals as opposed to an objective force apart from the system. The set of rules surrounding the algorithm meant that it was not possible to develop an algorithm fair to all as that was not the case of the English system in general. This goes against the idea that the process was objective as it was pursuing a precise political agenda.

Yet, while developing their system, they considered eleven possibilities based on their resources. They chose to eliminate some of those approaches on the basis of statistical analysis as those were not precise enough. They then identified two methods amongst the ones remaining that could be easily applied within exam boards. This was important as the main reason for this process was to allow boards to give certifications.

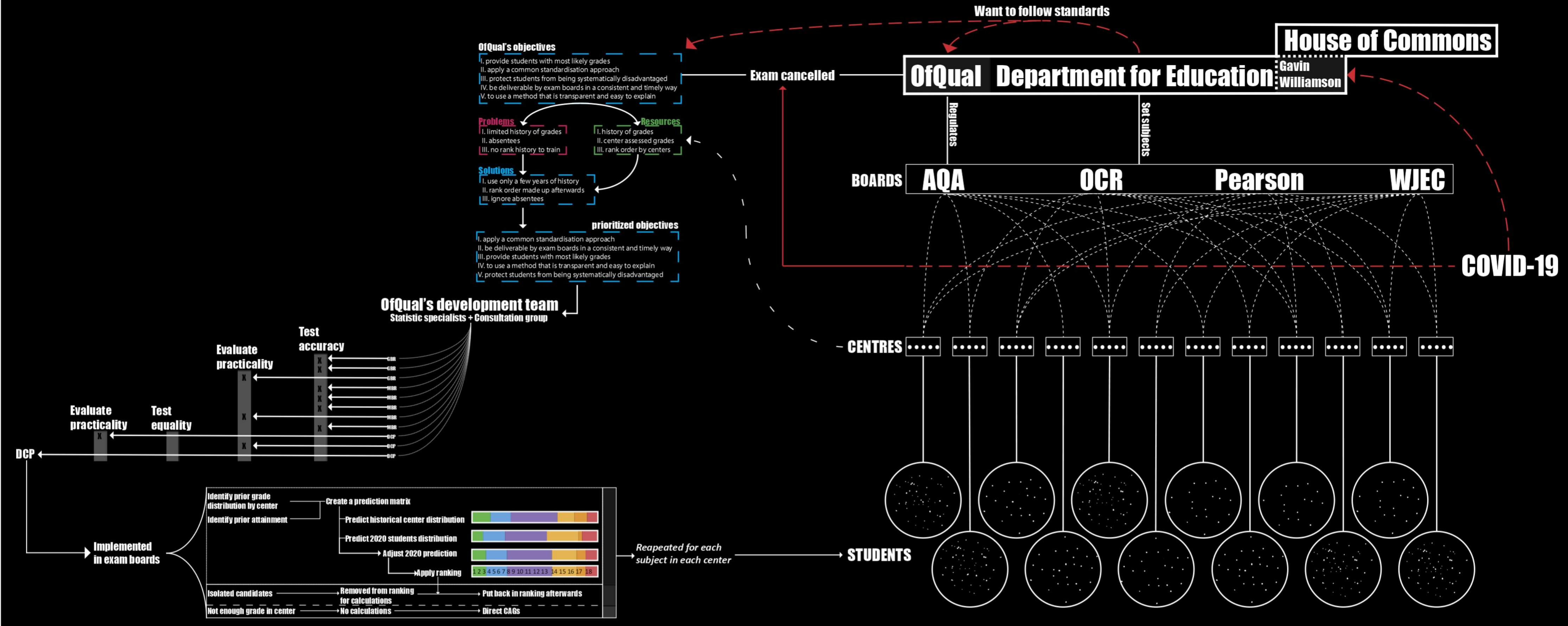
Those two methods were then tested to ensure that they did not enforce social discriminations. But as the test revealed that none of the approach was discriminatory, it showed the limits of auditing algorithms. Finally, the model used in the end was chosen on the basis of how easy it would be to fit in within exam boards.

Once again, we see here that the making of the algorithm is entangled within its societal environment as the practicality of its implementation played a major role in its development. Had the algorithm been developed in another context, the most practical system would have been different and would have produced other results.

The model was finally used across the educational system. But all the limitations and the ensuing negotiations meant that the so-called objective process put in place by OfQual was in fact very lenient towards students that are historically advantaged by the educational system. As it was strongly reliant on the prediction matrix identified from the history of each centre, it closely reproduced the general distribution of grades from previous years. All over, we saw how the making of the algorithm was influenced by the different limitations met and the advantages offered by its environment that nudged it in a particular direction. That goes to show how dependent the system was on its environment.

This goes in line with the social construction of technology approach by **Bijker** (1995) as he argued that technologies are negotiated amongst groups of actors. In our case, we saw that OfQual was in negotiation with the Department for Education to provide grades. Then, as those were being awarded, the previously muted group of students arose and went against the stabilisation of the technology by reopening negotiations.

Map of the implementation of the OfQual algorithm



Want to follow standards

House of Commons

OfQual

Department for Education

Gavin Williamson

Exam cancelled

COVID-19

Regulates

Set subjects

BOARDS AQA OCR Pearson WJEC

CENTRES

STUDENTS

OfQual's objectives
I. provide students with most likely grades
II. apply a common standardisation approach
III. protect students from being systematically disadvantaged
IV. be deliverable by exam boards in a consistent and timely way
V. to use a method that is transparent and easy to explain

Problems
I. limited history of grades
II. absentees
III. no rank history to train

Resources
I. history of grades
II. center assessed grades
III. rank order by centers

Solutions
I. use only a few years of history
II. rank order made up afterwards
III. ignore absentees

prioritized objectives
I. apply a common standardisation approach
II. be deliverable by exam boards in a consistent and timely way
III. provide students with most likely grades
IV. to use a method that is transparent and easy to explain
V. protect students from being systematically disadvantaged

OfQual's development team
Statistic specialists + Consultation group

Implemented in exam boards

Identify prior grade distribution by center

Identify prior attainment

Create a prediction matrix

Predict historical center distribution

Predict 2020 students distribution

Adjust 2020 prediction

Apply ranking

Isolated candidates → Removed from ranking for calculations → Put back in ranking afterwards

Not enough grade in center → No calculations → Direct CAGs

Reapeated for each subject in each center

CONCLUSION

Throughout our analysis, we studied how the making of the OfQual algorithm was entangled within its societal context. We have unpacked the algorithm and identified how it was in fact a perpetuation of the English educational system. It was a statistical model established by OfQual and enacted by the exam boards. As a direct consequence, the OfQual algorithm was more made to fit the educational system than the students.

We showed how the negotiations around the implementation of the algorithm were strongly imbued with the desire of maintaining standards by all means. We saw how the shift in OPP translated the usual examinations into the algorithm. We also identified how the algorithm was built on biases and negotiation and why it is misleading to present it as being an objective process. Indeed, it was strongly biased towards the idea that maintaining standards is both desirable and right. That allowed us to see why the algorithm was in fact indistinguishable from its environment.

As a result of the objectives they set, OfQual was unable to take into consideration the system as a whole and the consequences of implementing an opaque system to grade students. By putting in place their algorithm, OfQual also stripped the students from being able to try and prove their worth, as they were removed from the process of grading as a whole. This poses questions on the impact of putting in place such assemblages for the agency of the people being impacted by it. As the students were translated into data predicted by their teachers, they lost their agency in the final assemblage that was the algorithm.

Also, as Moses' bridge prevented poorer populations from attaining some public areas (**Winner**, 1986), the OfQual algorithm was stacked in advance with politics that led to discriminations against some categories of students. But unlike the bridge, the oppressive aspect of the algorithm came from the enactment of the artefact by OfQual and the boards.

While the code or the dataset could be isolated and studied in terms of what they allowed to do and the politics resulting from that, it is here the interaction between the different components of the algorithm that led to its discriminatory politics. In order to engage critically with the algorithm, it is thus capital to consider it as an assemblage of actants in interaction. Blaming the code or the dataset is thus not sufficient to efficiently uncover the politics of such systems.

As questions about grade standardisation start to appear within countries such as France (**Morin**, 2022), and as discussion about biases within algorithms are rising, it becomes more and more important to identify what algorithms are and how they are related to our own actions. During this work, we showed that it is not sufficient to point responsibility to the code or the dataset. The general network of relations leading to the implementation of discriminatory artefacts should also be taken in consideration when engaging with algorithms critically.

Moreover, the concept of biases within algorithms should be reconsidered. In the case of OfQual, we saw that the discriminatory results were not unintentional *per se*, it is not so much the code of the algorithm that was biased, but the assemblage of the algorithm as a whole. Further work could thus be made to study biases as being more than an undesirable effect in regard to the result of the algorithm, but also as a consequence of the implementation of such system in the first place.

Finally, while companies making algorithms such as **OpenAI** (2022) argue that their systems produce results that a human could also make, we saw that the outsourcing of decision making to an algorithm had dire consequences. Indeed, doing so, OfQual displaced the accountability of teachers to an indisputable black box. Displacing the power of humans to the code of automatic processes thus created a space where it was not possible to question the impending reproduction of the dominant system.

BIBLIOGRAPHY

Akrich, M., Callon, M., & Latour, B. (2006). *Sociologie de la traduction: Textes fondateurs*. 'Ecole des mines de Paris.

Algorithm, n. (n.d.). In *OED Online*. Oxford University Press. Retrieved 16 May 2022, from https://www.oed.com/view/Entry/4959

Benjamin, G. (2022). *#FuckTheAlgorithm: Algorithmic imaginaries and political resistance*.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.

Bijker, W. E. (1995). *Of bicycles, bakelites, and bulbs: Toward a theory of sociotechnical change*. MIT Press.

Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). *Multimodal datasets: Misogyny, pornography, and malignant stereotypes* (arXiv:2110.01963). arXiv. http://arxiv.org/abs/2110.01963

Bloor, D. (1976). *Knowledge and social imagery* (1st ed). University of Chicago Press.

Buolamwini, J., & Timnit, G. (n.d.). *Gender Shades*. Retrieved 30 June 2022, from http://gendershades.org/

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

Cardon, D. (2019). *Culture numérique*.

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

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. https://doi.org/10.1145/3531146.3533213

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Crawford, K., & Joler, V. (2018). *Anatomy of an AI System*. Anatomy of an AI System. Retrieved 16 May 2022, from http://www.anatomyof.ai

Deleuze, G., & Guattari, F. (1987). *A thousand plateaus: Capitalism and schizophrenia*. University of Minnesota Press.

Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society*, *3*(2), 205395171666512. https://doi.org/10.1177/2053951716665128

Educational Settings—Hansard—UK Parliament. (2020). Retrieved 19 June 2022, from https://hansard.parliament.uk//Commons/2020-03-18/debates/FCD4DEB2-86A8-4F95-8EB8-D0EF4C752D7D/EducationalSettings

Engels, F. (1872) "On Authority," in *The Marx-Engels Reader*, ed. 2, Robert Tucker (ed.) (NewYork:W. W. Norton, 1978),731.

Hao, K. (n.d.). *The UK exam debacle reminds us that algorithms can't fix broken systems*. MIT Technology Review. Retrieved 16 October 2020, from https://www.technologyreview.com/2020/08/20/1007502/uk-exam-algorithm-cant-fix-broken-system/

Heikkila, M. (2022, March 29). *Dutch scandal serves as a warning for Europe over risks of using algorithms*. POLITICO. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

Jasanoff, S., & Kim, S.-H. (2009). Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea. *Minerva*, *47*(2), 119–146. https://doi.org/10.1007/s11024-009-9124-4

Joler, V., & Pasquinelli, M. (n.d.). *The Nooscope Manifested*. Retrieved 23 May 2022, from https://nooscope.ai/

Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S. G., Noble, S. U., & Shestakofsky, B. (2021). Toward a Sociology of Artificial Intelligence: A Call for Research on Inequalities and Structural Change. *Socius: Sociological Research for a Dynamic World*, *7*, 237802312199958. https://doi.org/10.1177/2378023121999581

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

Knight, W. (n.d.). The Apple Card Didn't 'See' Gender—And That's the Problem. *Wired*. Retrieved 29 June 2022, from https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Latour, B., Mauguin, P., & Teil, G. (1992). A Note on Socio-Technical Graphs. *Social Studies of Science*, *22*(1), 33–57. https://doi.org/10.1177/0306312792022001002

Louridas, P. (2020). *Algorithms*. The MIT Press.

Mac, R. (2021, September 3). Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men. *The New York Times*. https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html

Mau, S. (2019). *The metric society: On the quantification of the social*. Polity Press.

Morin, V. (2022, June 23). Le baccalauréat au défi d'une inflation « irrépressible » des notes. *Le Monde.fr*. https://www.lemonde.fr/societe/article/2022/06/23/le-baccalaureat-au-defi-d-une-inflation-irrepressible-des-notes_6131751_3224.html

Newton, P., Goldstein, H., Patrick, H., & Tymms, P. (2008). *Techniques for monitoring the comparability of examination standards*.

OfQual. (2020a). *Information for heads of centres.* OfQual.

OfQual. (2020b). *Consultation on specified general qualifications*. OfQual.

OfQual. (2020c). *Awarding GCSE, AS, A level, advanced extension awards and extended project qualifications in summer 2020: Interim report* (p. 319). OfQual.

Ofqual. (2020d, July 21). *Calculating grades in GCSE, AS and A levels summer 2020*. https://www.youtube.com/watch?v=EX5STb0qbGI

OfQual. (2020e). *Standardisation of grades in general qualifications in summer 2020: outliers*. OfQual.

OfQual. (2021). *Summer 2021 student-level equalities analysis—GCSE and A level*. GOV.UK. https://www.gov.uk/government/publications/analysis-of-results-a-levels-and-gcses-summer-2021/summer-2021-student-level-equalities-analysis-gcse-and-a-level

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First edition). Crown.

OpenAI. (2022). *Openai/dalle-2-preview*. OpenAI. https://github.com/openai/dalle-2-preview/blob/33c8f159b27b66b6129cf85e8eb1b16701b8099e/system-card.md (Original work published 2022)

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

Panch, T., Mattie, H., & Atun, R. (2019). Artificial intelligence and algorithmic bias: Implications for health systems. *Journal of Global Health*, *9*(2), 010318. https://doi.org/10.7189/jogh.09.020318

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *Journal of Global Health*, *8*(2), 020303. https://doi.org/10.7189/jogh.08.020303

Reshuffle: Boris Johnson fires Gavin Williamson as he rings cabinet changes. (2021, September 15). *BBC News*. https://www.bbc.com/news/uk-politics-58571935

Review into bias in algorithmic decision-making. (n.d.). GOV.UK. Retrieved 30 June 2022, from https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2), 205395171773810. https://doi.org/10.1177/2053951717738104

Seurat, C., Tari, T., & Latour, B. (2021). *Controverses, mode d'emploi*. Sciences po, les presses.

Summer 2020 code used to grade qualifications. (2022). [R]. Ofqual. https://github.com/OfqualGovUK/Summer-2020-code-used-to-grade-qualifications (Original work published 2020)

Williamson, G. (2020, March 31). *DIRECTION UNDER S 129(6) OF THE APPRENTICESHIPS, SKILLS, CHILDREN AND LEARNING ACT 2009*. 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.