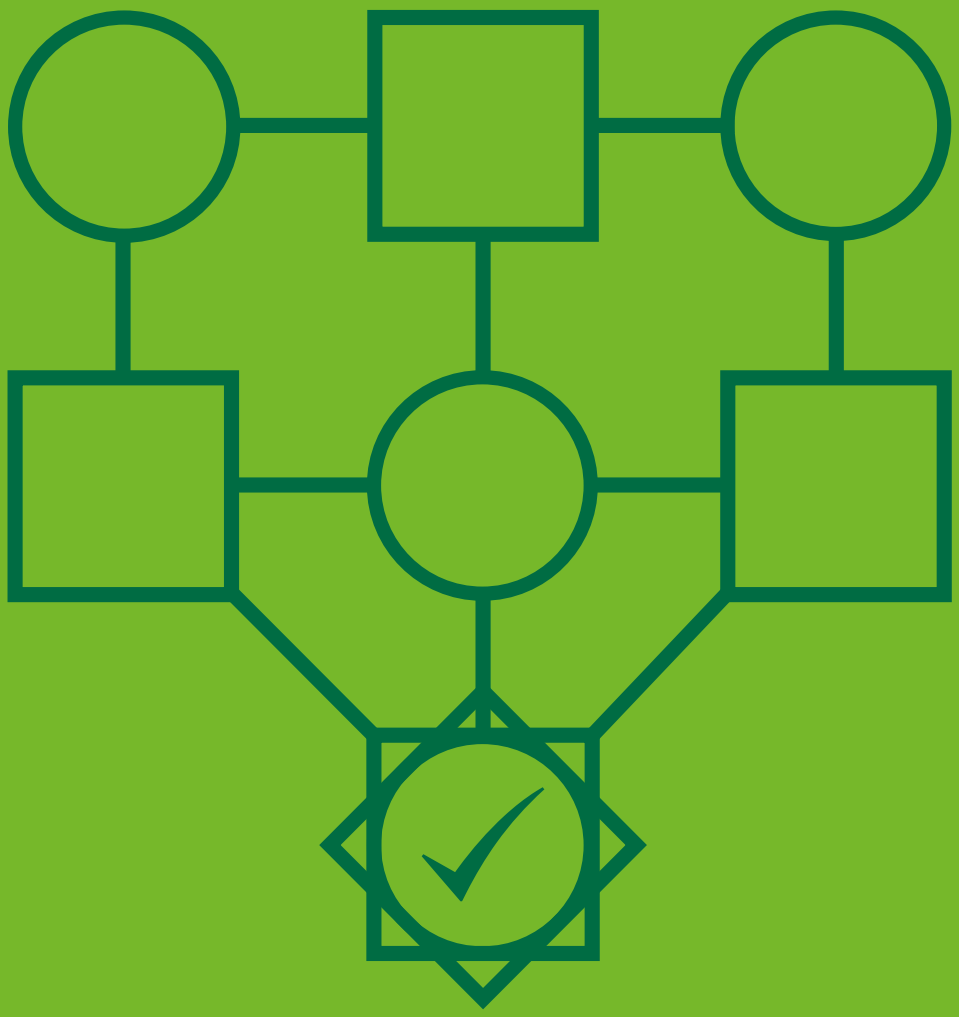
bcs
The Chartered Institute for IT

2020 | REPORT

# THE EXAM QUESTION:

## HOW DO WE MAKE ALGORITHMS DO THE RIGHT THING?



Making IT good for society

# THE EXAM QUESTION:

## HOW DO WE MAKE ALGORITHMS DO THE RIGHT THING?

## CONTENTS

1 THIS REPORT IN A SENTENCE
2 EXECUTIVE SUMMARY
3 THE USE OF ALGORITHMS TO ESTIMATE EXAM GRADES AS A CASE STUDY
- 3.1 What the Ofqual 'algorithm' does
- 3.2 Lessons on public policy standards from the Ofqual estimated grade system
- 3.3 Openness
- 3.4 Accountability
- 3.5 Objectivity

4 KEY STAGES OF DESIGN AND DEVELOPMENT FOR INFORMATION SYSTEMS
5 PROFESSIONALISING DEVELOPMENT OF INFORMATION SYSTEMS THAT ARE BASED ON ALGORITHMS
6 GLOSSARY
7 WHO WE ARE - BCS, THE CHARTERED INSTITUTE FOR IT

# 1. THIS REPORT IN A SENTENCE

An analysis of the Ofqual exam grade awarding algorithm to identify how principles of openness, accountability and objectivity can be embedded in algorithms that make high-stakes judgements about people.

# 2. EXECUTIVE SUMMARY

Recently people have started referring to algorithms as 'prejudice engines' that should never be used in high-stakes public policy matters[1], which is highly misleading and has the potential to damage public trust and confidence in algorithms as a fundamental building block of information systems. Information systems that rely on algorithms can be a force for good and a lot of the time they are, but we've woken up to the fact that in high-stakes situations it's very hard to make them work as we intend them to. It's our assertion that rather than focus on algorithms per se, we should be looking at how we can professionalise data science so that the UK has the most trusted, ethical and sought-after data science teams in the world, building on the findings of the Royal Society report on data science skills[2].

The overarching issue that should be a matter of urgent public debate is the way we have been formulating public policy and delivering public services through information systems that use algorithms to make high-stakes judgements about individuals based on data that is subjective, uncertain in its quality and provenance and partial. The Committee[3] on Standards in Public Life (CSPL) have highlighted that where automated information systems are intended to provide public benefit they have to meet the same standards that people in public office have to meet, which especially includes adhering to the principles of openness, accountability and objectivity[4].

A significant challenge is embedding those principles throughout design and development stages and ensuring the interdisciplinary teams who manage and maintain information systems have governance processes capable of upholding those principles at the system level. That includes oversight of the way algorithms are used, including identifying unintended consequences, and the capability to remedy harm that might be caused to an individual when something goes wrong. Another significant challenge is getting us, the public, to understand what it means for an information system to be open, accountable and objective and how to gain our trust that a system is effective at upholding those principles.

This paper examines how in future we can design automated information systems to benefit all of us through better openness, accountability and objectivity and how to establish trust that systems deliver against those principles. As a mechanism for identifying and exploring the issues in greater detail we examine the recent use of algorithms as part of the Ofqual grade estimation system to award A-level and GCSE grades in England[5], which caused so much public condemnation that government policy had to be reversed.

1 https://twitter.com/TimandraHarknes/status/1295606814629912582
2 https://royalsociety.org/-/media/policy/projects/dynamics-of-data-science/dynamics-of-data-science-skills-report.pdf?la=en-GB&hash=212DAE7D599B0A48687B372C90DC3FEA
3 CSPL advises the Prime Minister on ethical standards across the whole of public life in England, https://www.gov.uk/government/organisations/the-committee-on-standards-in-public-life
4 https://cspl.blog.gov.uk/2020/08/19/decision-making-by-algorithm-must-meet-nolans-tests/
5 https://www.gov.uk/government/news/guide-to-as-and-a-level-results-for-england-2020

We recommend the following steps be taken in order to help restore public trust in the openness, accountability and objectivity of information systems that rely on algorithms and data:

- That Government support the ongoing collaboration between the Royal Statistical Society, BCS, the Operational Research Society, the Royal Academy of Engineering, the National Physical Laboratory, the Royal Society and the Institute of Mathematics and its Applications, to professionalise data science.
- Learned societies and Government should work collaboratively to establish forums and communities of practise to facilitate policy makers appreciation of the capabilities, opportunities and risks of data driven automation of policy formulation and implementation.
- Government takes a strong leadership role to ensure already established good ethical and professional practice in algorithm design, development and testing become ubiquitous across all of government and industry.
- All information systems that rely on algorithms and data that are used in any public policy must go through an impact assessment documented publicly, undertaken by independent experts against an appropriate ethical framework[6] before going live, which are made publicly available. A review of impact assessment in this area should be undertaken to understand best practice in transparency and impact.

The current exam grading situation[7] should not be attributed to any single government department or office. In our view it is a consequence of fundamentally repurposing government organisations in a time of national crisis to build information systems that rely on algorithms and data to work properly, and for which they have no prior experience in developing. By systematically adopting ethical and professional standards and good professional practice they will be equipped to deliver innovative information systems in future that meet public expectations of fairness and value.

"THIS PAPER EXAMINES HOW IN FUTURE WE CAN DESIGN AUTOMATED INFORMATION SYSTEMS TO BENEFIT ALL OF US THROUGH BETTER OPENNESS, ACCOUNTABILITY AND OBJECTIVITY AND HOW TO ESTABLISH TRUST THAT SYSTEMS DELIVER AGAINST THOSE PRINCIPLES."

6 Such as the IBM data analytics and big data ethical framework https://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD%26A.pdf

7 https://www.bbc.co.uk/news/technology-53836453 "Ofqual's 'cheating' algorithm under review"

# 3. THE USE OF ALGORITHMS TO ESTIMATE EXAM GRADES AS A CASE STUDY

In the summer of 2020 A-level and GCSE examinations in England were cancelled because of the COVID-19 pandemic. In their place students were awarded grades based on an estimation calculated by Ofqual using predictions from teachers together with historical school performance data. The estimation was widely reported as produced by an 'algorithm' and has been widely criticised as unfair and unjust. The algorithm is one component of an information system used to automate the award of grades to students when no examinations have taken place. That information system is the result of many design choices that are the consequence of decisions about public policy, governance structures, data model selection, the choice of a criteria for allowing people to appeal against awarded grades, through to deciding what it means for the system to be fit for purpose and how to test it.

## 3.1 WHAT THE OFQUAL 'ALGORITHM' DOES

This section briefly summarises at a very high level how the grade estimation system developed by Ofqual calculated the grades awarded to students in those circumstances that were viewed as particularly controversial, which is based on Ofqual's official report[8] and informed by a variety of independent opinions[9],[10]¬. We are not attempting to give an exact description of the algorithm here, rather we are trying to convey in a way that is accessible to a general reader how the algorithm was applied in cases that were seen by many as highly controversial.

Ofqual based their calculations for awarding grades on a model comprised of:

- the historical performance of students that had been entered for each subject in each school
- prior-attainment relating to previous students across all schools from historical data-sets
- prior-attainment of the students in each school who were due to sit exams this summer
- teacher assessed grades and how they ranked students compared to their peers

Ofqual knew that much of the data available to them contained a lot of uncertainty and in some cases was incomplete, and they did everything they could to take account of those complications. In particular Ofqual regarded predicted grades from teachers as uncertain and variable, whereas they had greater confidence in the ability of teachers to correctly rank students' performance in relation to other students in their peer group.

Without going into the specific details, what appears to have happened for many schools (but by no means in all cases) is roughly the following:

- The expected spread of grades that each school cohort, in each subject, should get was calculated based on various types of historical data i.e. how many students were expected to achieve grades A*, A, B, C, etc at A-level in each subject offered by a school/college, and a similar calculation was done for how many students were expected to achieve grades 9, 8, 7, 6, etc at GCSE in each subject.

8 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/909368/6656-1_Awarding_GCSE__AS__A_level__advanced_extension_awards_and_extended_project_qualifications_in_summer_2020_-_interim_report.pdf
9 https://unherd.com/2020/08/how-ofqual-failed-the-algorithm-test/
10 https://thaines.com/

› The way students were ranked in comparison to each other by their teachers was then used to give each of this year's students a specific grade, from the calculated historical spread of grades i.e. the student with the lowest ranking was allocated the lowest grade in the historical spread of grades, and the highest ranked student was allocated the highest grade in the historical spread of grades. The other students were then allocated grades from the historical spread in a way that kept their relative ranking to each other and which was informed by other historical data on students' prior-attainment.

Under these circumstances the Ofqual estimated grade might be very different from the teacher's predicted grade for a student. When Ofqual felt the cohort of students taking a subject was small it instead relied far more on teacher's predicted grades rather than the ranking or historical data for that subject. That also led to controversy because often independent schools offer a greater diversity of subjects resulting in small cohorts for those subjects.

Consequently, a large number of students were reported in the press as being awarded grades that appeared to have no relation to the grades predicted by their teachers (which is distinct from the relative ranking of students by teachers). The resulting public outcry and condemnation ultimately led to a reversal of government policy, which meant the Ofqual estimated grades system, in its original form, was abandoned in favour of one that effectively awarded grades based on teacher predicted grades. This is also controversial since it has led to significant grade inflation, which is one of the key policy issues that Ofqual were instructed to avoid. The Chief of Ofqual had to resign as a result of the public outcry over the estimated grades initially awarded by Ofqual[11].

## 3.2 LESSONS ON PUBLIC POLICY STANDARDS FROM THE OFQUAL ESTIMATED GRADE SYSTEM

Current publicly available information about the way DfE and Ofqual have chosen to automate the estimation of exam grades for students in England, have highlighted a number of important issues for designing information systems, that depend on algorithms that are intended to be open, accountable and objective. This section explores what those are.

This section is split into subsections covering openness, accountability and objectivity.

## 3.3 OPENNESS

DfE and Ofqual conducted extensive consultations with a wide range of stakeholders about whether calculated grade estimation should be used as a replacement for exams, and what type of data should be used as the basis for the calculations. Consultations covered questions about whether using teacher predicted grades and historical data on school performance were the right data to use, and also covered questions about the need to maintain standards. The BCS school curriculum and assessment committee were one of the stakeholders who were consulted[12]. One of the purposes of the consultation was to determine types of data that were appropriate to use, and principles, such as maintaining standards, that would form the key components of the high level model the grade estimation system was then based on.

Stakeholders were not consulted on several other important issues such as what was the best method for using the collated data to estimate grades or how best to maintain standards. They were also not consulted on how to combine a teacher's predicted grade with historical data on school performance to decide what grade to award or how that would be done in a way to guarantee standards were maintained, i.e. how to structure the algorithm that decides the grades to award. Lastly, stakeholders were not consulted on what it means for the system to be fair and reasonable as well as how best to maintain standards. Had these questions been included in the consultation process it is likely issues would have been highlighted early on.

11 https://www.bbc.co.uk/news/education-53909487
12 https://www.bcs.org/more/about-us/people-and-governance/our-boards-and-committees/school-curriculum-and-assessment-committee/policy-papers-and-submissions/

**Conclusion:**

Openness means being open about what data will be used, the provenance of that data, how it will be used and what criteria will be used to determine whether the resultant information system is fit for purpose.

## 3.4 ACCOUNTABILITY

Ofqual is accountable to Parliament, the DfE is accountable to the elected government of the day through the Secretary of State for Education. Ofqual is not accountable to the Secretary of State for Education. The Ofqual board is the legal authority responsible for the regulation of qualifications, examinations and assessments.

With hindsight it seems reasonable to say that more robust governance mechanisms across Parliament, DfE and the Ofqual board were needed in order for them to jointly oversee the design and development of the exam grade estimation system. Due to the urgency of the situation and the unprecedented nature of the problem it was essential these stakeholders were able to work collaboratively during the design and development process to identify and resolve issues as they occurred and to keep each other fully informed about potential risks as they became apparent.

**Conclusion:**

It is essential to develop effective mechanisms for the joint governance of the design and development of information systems right at the start, including all the stakeholders who need to have oversight, those that need to make it work, as well as those responsible for its success.

## 3.5 OBJECTIVITY

The Commons Select Committee on Education raised a number of concerns[13] before the estimated grading systems were used to award grades to students, such as:

- *"We believe that modifications to assessments will lead to erosion of standards, and that the 2020 cohort of exam-takers could be disadvantaged by a perception that their exams were not as rigorous as those taken by other cohorts."*
- *"Given the potential risks of bias in calculated grades, it is clear that standardisation will be a crucial part of ensuring fairness. We are extremely concerned that Ofqual's standardisation model does not appear to include any mechanism to identify whether groups such as BAME pupils, FSM eligible pupils, children looked after, and pupils with SEND have been systematically disadvantaged by calculated grades."*

Ofqual addressed concerns such as those of the Select Committee by extensive and robust checks on the overall *distribution* of grades across different groups of students based on their characteristics and demonstrated that *statistically* they would get grades that met these expectations. However, that is completely different to checking if there are going to be a large *number* of students from a particular group that will be awarded grades that a reasonable member of the public would regard as *unjust*.

**Minister of State Nick Gibb has said[14]:** *"It was not intended that a young person who had worked diligently for two years on their A-levels and was expecting an A and two Bs or three As, and turned up at school to collect their grades and they were three Ds."* This implies the minister believed individuals should feel they had been treated fairly with respect to the grades predicted by their teachers, which is different to fairness in the sense the overall grades for cohorts of students with similar characteristics are statistically similar to previous years.

**Ofqual's top level requirements as they saw them are explained by Roger Taylor, Chairman of Ofqual:** *"There was no easy solution to the problem of awarding exam results when no exams have taken place. Ofqual was asked by the Secretary of State to develop a system for awarding calculated grades, which maintained standards and ensured that grades were awarded broadly in line with previous years. Our goal has always been to protect the trust that the public rightly has in educational qualifications."*

13 https://publications.parliament.uk/pa/cm5801/cmselect/cmeduc/617/61708.htm
14 https://www.theguardian.com/education/2020/aug/20/minister-nick-gibb-admits-he-was-warned-about-concerns-over-exams-algorithm

If a stated policy objective had been that any reasonable member of the public would believe every student is treated as fairly as possible, then how that is achieved objectively would have been a topic that could be resolved by DfE, Ofqual and Parliament jointly. In other words, it is important to establish right at the start what it will mean for the public to have trust in an information system based on algorithms and to decide how that can be done objectively.

**Conclusion:**

Be clear what is intended to be achieved at an individual level for all those affected by an information system, how that is going to be objectively measured, and be clear what it will mean for individuals to have trust in the system.

"IT WAS NOT INTENDED THAT A YOUNG PERSON WHO HAD WORKED DILIGENTLY FOR TWO YEARS ON THEIR A-LEVELS AND WAS EXPECTING AN A AND TWO BS OR THREE As, AND TURNED UP AT SCHOOL TO COLLECT THEIR GRADES AND THEY WERE THREE Ds."

**NICK GIBB**

# 4. KEY STAGES OF DESIGN AND DEVELOPMENT FOR INFORMATION SYSTEMS

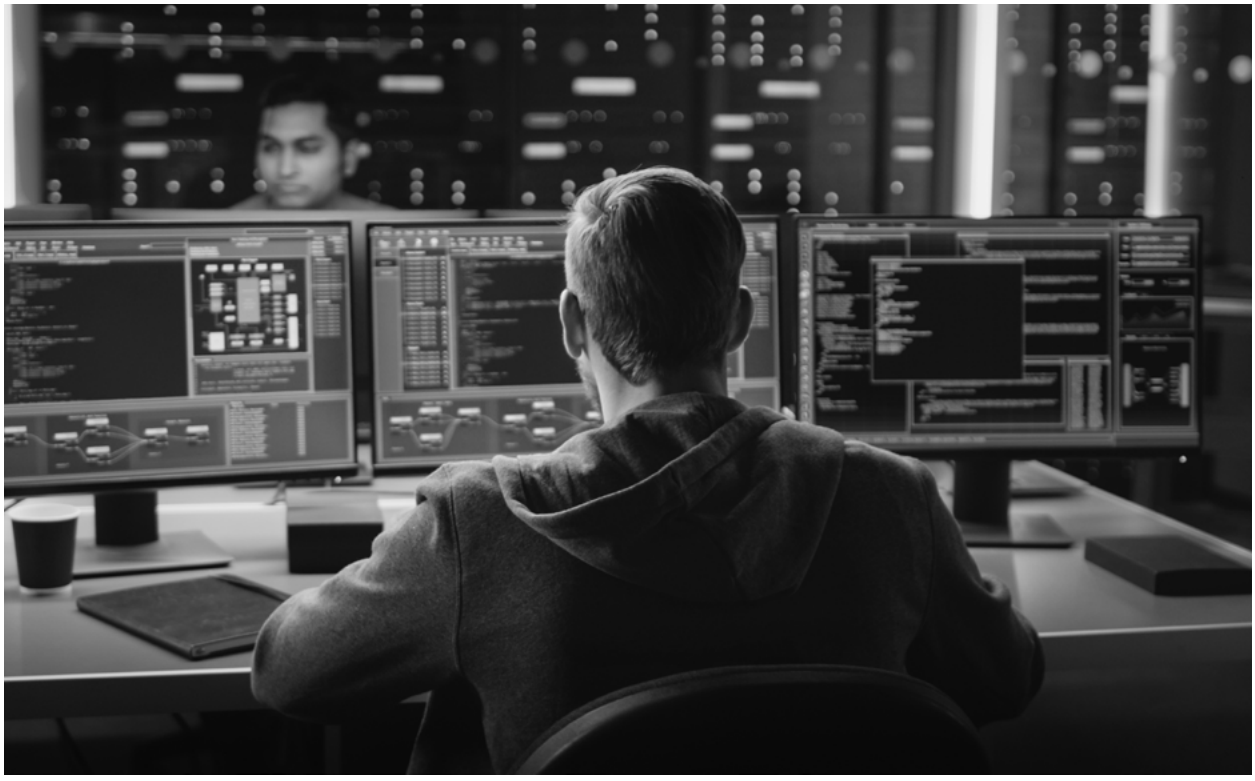
This section briefly summarises the different stages that affect the design and development of an information system meant to support public policy objectives. This is intended to help understand how any algorithm used in an information system is a consequence of multiple layers of choices made by different stakeholders at different times.

- **Policy objectives** – right at the start policy makers need to frame the outcomes they want in a way the public can understand and in a way they can be effectively consulted on. Choices at this point about what data will be needed and how it will be used to deliver policy objectives will have a huge influence on what information system is produced and any supporting algorithms.
- **Ownership** – at this stage choices have to be made about which organisations are responsible for delivering the information system, and who needs to have oversight of its development. This stage is key to ensuring stakeholders are able to check the final information system is going to deliver the right outcomes. That includes testing the system against the principles of openness, accountability and objectivity. Choices about how to implement those principles will affect the design of the information system and the underpinning algorithms.
- **The model** – deciding on a data model that will lead to the intended outcomes is the result of choices that to some degree are subjective. All of these choices affect what the information system can and can't do and how well it does them.
- **Gathering and processing data** – identifying and capturing the appropriate quality data based on the chosen data model and making sure it can be turned into a form that is fit for purpose is challenging and involves many choices. It is critical that an evaluation of the processed data is carried out prior to its use in informing judgements and it is equally important that those going to be affected by the data have sight of the evaluation. Once again choices about how to evaluate the data will affect how the data is used, which again means more choices about any final algorithms.
- **The algorithm** – only after choices about policy objectives, ownership, data models and data gathering are made can a set of algorithms be developed that collectively automate those judgements the data is fit to be used for. Algorithms have consequences beyond the creation of a computer program, they constrain how objectively stakeholders can interrogate information that has been processed, which shapes the way people choose to present outputs to others when explaining things like fairness, and greatly affects any appeals process. Yet again there are more choices about the exact steps an algorithm follows and what counts as an exception the algorithm should escalate for someone or something else to handle through an appeals process.

- **Testing is vitally important** – this will determine if the system is good enough. At several points in system development thorough testing will be required. This will include testing standalone parts of the system, how well those distinct elements work when integrated into a bigger component, and how well the final system achieves what it was intended for, including testing the system with those it affects as well as those who use it. Deciding what will need to be tested and how to test it will influence the design of a system. Testing average outcomes over populations leads to one set of design choices, testing outcomes on an individual basis leads to different design choices.

All of the above stages require sound ethical judgement to make the best choices possible. They involve many different stakeholders who need the right governance mechanisms to work in close collaboration both quickly and effectively. This is not straightforward, which is why much effort has gone into establishing good practice and ethical and professional standards in the IT profession over decades.

"THIS ALGORITHM IS MORE THAN A PIECE OF COMPUTER PROGRAMMING, IT IS ALSO THE INTERFACE TO STAKEHOLDERS, THE WAY THE INFORMATION IS PRESENTED, AND ANY APPEAL PROCESS."



# 5. PROFESSIONALISING DEVELOPMENT OF INFORMATION SYSTEMS THAT ARE BASED ON ALGORITHMS

What are the challenges for policy makers in commissioning the development of information systems that meet public expectations of ethical outcomes?

Currently there is a lack of clearly defined practical mechanisms that practitioners need to be equipped with in order to thoughtfully embed ethical principles throughout the lifecycle of algorithmic systems used in high-stakes policy implementation, from design to deployment and across business management processes. There appears to be a lack of professional good practice and professional standards employed around the development and implementation of information systems. Our views at the BCS are consistent with the findings of the Royal Society report on Data Science Skills[6].

One reason the problem of ethical algorithmic systems practice is difficult is because it is much more than just understanding ethical principles and more than just learning how to code technical solutions to ethical challenges into systems. Ethical practice includes being able to thoughtfully interact with organisational managers, data scientists, systems engineers, legal departments, and policy makers etc, to help them understand how ethical concerns need to be addressed across an organisation. Another difficulty stems from the fact that algorithms are being applied in real-time policy contexts in novel ways that make it difficult to uncover how and why poorly constructed decision-making models are generated and how to put in place effective governance to remedy issues as soon as they occur.

An important principle is to treat any automated information system that relies on statistically based best judgement learnt from acquired data in order to make decisions in real time, as a high risk algorithmic system. Note there are cases where the algorithm is not necessarily a risk in itself. Such as systems that make decisions based solely on deterministic rules, or where decisions are based on statistical analysis, and where outputs can be scrutinised off-line by people to validate them before they are implemented (which historically has been the case in Business Analysis systems). Importantly high risk systems include Machine Learning systems that are able to learn latent variables[15] from large data sets, that are beyond human scrutiny, and that are used to create decision making models that discriminate between people because of those latent variables.

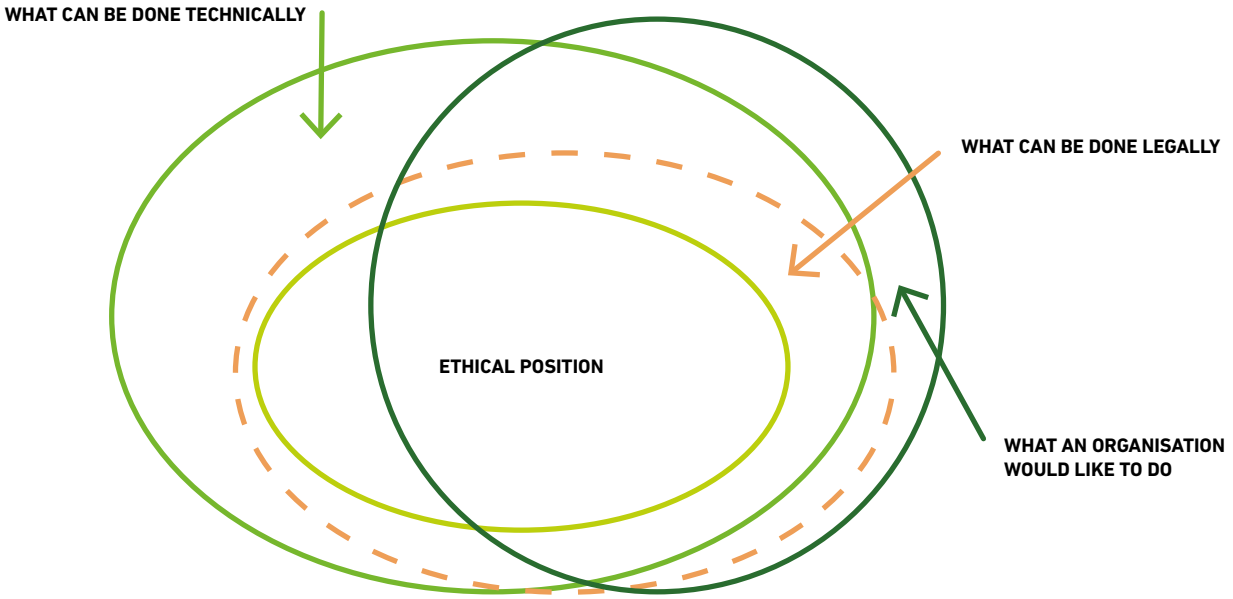
More precisely then, especially high risk algorithmic systems include:

- automated systems that must process data streams in real-time;
- that use probabilistic self-learning algorithms to inform decisions that will have significant consequences for people;
- used where it is difficult to uncover how decisions are derived;
- used where contestability of decision is not deterministic and
- ultimately decisions rely on best judgment that requires understanding of the broader context

Mitigating the risks caused by such systems in policy formulation or implementation requires understanding all the organisational business practices and how interdisciplinary teams work together across policy boundaries. Figure 1 illustrates overlapping organisational constraints that can lead to difficult ethical choices[16].

Figure 1:

OVERLAPPING CONSTRAINTS FROM IBM'S ETHICS OF BIG DATA AND ANALYTICS REPORT BY MANDY CHESSELL




Source: IBM Ethics for big data and analytics white-paper 2014

All of this complexity and interdependencies reinforces the need to professionalise information systems design and development across government. Fortunately work on professionalising these practices has already begun[17], with the Royal Statistical Society working in partnership with BCS, the Operational Research Society (ORS), the Royal Academy of Engineering (RAEng), the National Physical Laboratory (NPL), the Royal Society and the IMA (The Institute of Mathematics and its Applications), to collaboratively shape and develop the data science profession. We recommend Government support and join this collaboration, ensuring that it enthusiastically adopts the professional standards and practise that are developed by this partnership.

15 Which means variables that are not manually specified by designers and are not explicitly present in the model, nor present as inputs or outputs of the model.
16 https://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD&A.pdf
17 https://rss.org.uk/news-publication/news-publications/2020/general-news/professional-standards-to-be-set-for-data-science/

# 6. GLOSSARY

| | |
|---|---|
| **BAME** | Black and minority ethnicity |
| **BCS** | British Computer Society |
| **CSPL** | Committee on Standards in Public Life |
| **DfE** | Department for Education |
| **FSM** | Free school meals |
| **GCSE** | General Certificate of Secondary Education |
| **IMA** | Institute of Mathematics and its Applications |
| **NPL** | National Physics Laboratory |
| **Ofqual** | Office for Qualifications |
| **RAEng** | Royal Academy of Engineering |
| **RSS** | Royal Statistical Society |
| **SEND** | Special educational needs and disabilities |

# 7. WHO WE ARE - BCS, THE CHARTERED INSTITUTE FOR IT

BCS is the UK's Chartered Institute for IT. The purpose of BCS as defined by its Royal Charter is to promote and advance the education and practice of computing for the benefit of the public.

We bring together industry, academics, practitioners and government to share knowledge, promote new thinking, inform the design of new curricula, shape public policy and inform the public.

As the professional membership and accreditation body for IT, we serve over 60,000 members including practitioners, businesses, academics and students, in the UK and internationally.

We also accredit the computing degree courses in ninety eight universities around the UK. As a leading IT qualification body, we offer a range of widely recognised professional and end-user qualifications.

For further information please contact:

**BCS**
The Chartered Institute for IT
3 Newbridge Square
Swindon
SN1 1BY

**T** +44 (0)1793 417 417

www.bcs.org