AI and Social Justice

Notes of talk at the AI Summit New York, 8th Dec. 2021

https://www.alandix.com/academic/talks/AI-Summit-NY-2021-AISJ/
https://alandix.com/ai4sj/

Clara Crivellaro
Open Lab, Newcastle University, UK

Alan Dix
Computational Foundry, Swansea University, Wales, UK

Abstract

AI and in particular large data machine learning are transforming many areas of society including healthcare, education, and finance. At their best these offer the potential to improve society, for example, finding new pharmaceuticals. However, they may also reproduce or reinforce existing divisions and inequalities as well as creating new problems. This has been evident in high-profile news items such as the case of racial discrimination in facial recognition systems used in policing. However, some of the deepest problems in unequal access to technology are still to be fully felt.

Fortunately, AI can also be used positively to address issues of social injustice, for example human-rights organisations scanning public domain images for evidence of abuse, or software using adversarial techniques to reduce bias in training data. At best some companies and institutions are addressing these issues proactively, seeking ways to ensure they prevent or detect problems before they happen, for others this may be a rear-guard action to fix problems that have already emerged.

In this talk we will present a high-level landscape of the ways on which AI interacts with social justice and illustrate this through examples so that we can take positive action for a fairer world.

Keywords: artificial intelligence, social justice, bias, digital inclusion, digital economy, design, machine learning

Introduction

Overview

AI is radically transforming how we live, how we make decisions, the way we work, how we relate to one another, and how services in our cities are run in all domains of public life—from high stake domains such as legal systems, health-care, and education to more mundane aspects such as how we go about shopping or listening to music. Further, AI as an industrial complex is inevitably contributing to the shaping of our planetary ecosystem – thus its role should be considered in our concerted efforts to respond to our current ecological crisis.

Despite rightful enthusiasm for the potential of AI capabilities to enable more efficiency and support ways to expose patterns in large data-sets and information for more meaningful and informed decision-making—these transformations are not benefiting everyone in our increasingly 'digital-by-default' society. Rather AI has received quite a lot of bad press in recent years, where it is understood to be reproducing socio-economic inequality and even creating new forms of social division and exclusion. Yet, statistics and data science (thus AI) can have a role in emancipatory, human rights and social justice projects.

In the following, we draw from different examples to chart how we can move from avoiding harm (beyond racial and gender bias in AI) to supporting positive action while considering wider issues of access. We hope our reflections can contribute to the ways we can conceive of our responsibilities for socio-ecological justice with and through AI.

Background

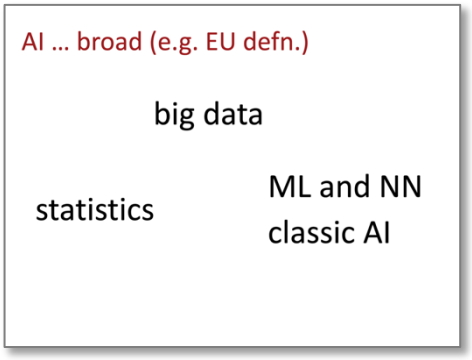
This talk arose in part due to EPSRC funded Network+ project Not Equal that Clara and Alan have been part of for the past three years. This network funded many smaller projects within three main themes: algorithmic social justice, digital security for all, and fairer futures for business and workforce. The topic of this talk relates especially to the first of these themes. In particular, we considered how little existing (socio-material) inequalities and social justice are at the centre of our concern when we design and deploy AI.

In addition, Clara and Alan are working on a book together "AI for Social Justice" (https://alandix.com/ai4sj/), which will extend many of the areas discussed in these talk notes.

What is AI?

A common definition of Artificial intelligence is any algorithm or computer system which performs a task that, if a human were to have performed it, would be regarded as intelligent. This includes

more logical or deductive tasks such as chess or mathematics and also more subconscious behaviour such as recognising a face in a photograph. In the early days of AI these types of tasks were coded using parallels with human cognition both high-level (such as logics) and low-level (such as neural networks), but it also become apparent some years ago that tasks that include learning from sheer volume of data can appear to behave in intelligent ways [HN09].



The current European Union definition of AI [EC21] includes all three forms of reasoning:

> *(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;*
>
> *(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;*
>
> *(c) Statistical approaches, Bayesian estimation, search, and optimization methods.*
>
> Annex I, Artificial Intelligence Techniques and Approaches [EC21]

This definition is still being debated, however here we will not distinguish these different technologies as all tend to led to similar issues for social justice.

What is Social Justice?

It is even harder to define social justice than AI as the term means so many different things to different people. Social justice is distinguished from personal ethics or individual justice as it looks at the overall distribution of power and resources within society and the impacts of institutions for groups as a whole. Whether you give money to support a food bank is a personal ethical decision, whereas the issues as to whether people should be in a state of poverty that necessitates food banks is a question of social justice. Beyond this, things get more complex. There are attempts in philosophy to tie it down more precisely, for example Rawls concept of a 'veil of ignorance' leading to distributive justice [Ra71], but these are themselves heavily debated.

Indeed, social justice has been variously defined throughout history. From its association with the law and the language of rights, where social justice finds gaps between definitions and interpretation of rights and their practical application; to collective efforts to redefine what may or ought to be considered 'just' or 'right' and for whom– reshaping collective understanding of social issues and initiating processes of transformation as to the ways we should respond to these issues. In this sense, social justice is better understood not as something static but rather as about ongoing practical achievements and the result of distributed, and collective efforts.

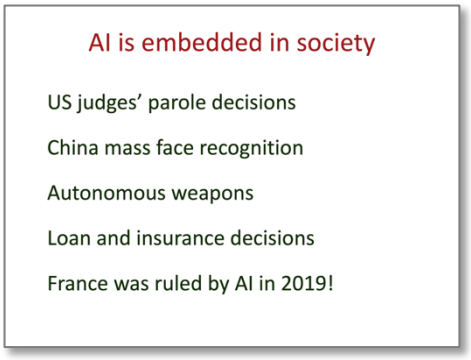
Knowing that readers differ in their political and social viewpoints, we will try to present the issues relating to the interplay between AI and social justice in ways that you can work out for yourselves how they might influence your actions in terms of the use and applications of AI. However, inevitably our own views colour the kinds of issues we discuss and the assumptions we make.

This said, there are ideas that are widely shared. We expect the institutions of society to operate impartially across different groups. When AI systems have been shown to embody racial bias, this has led to near universal condemnation. In retrospect, we can look back at campaigns for women's voting rights and racial equality as embodying principles that few would challenge now, albeit these were less clear to many at the time. Today as digital technology changes the landscape, we are at a similar point of flux. In twenty years, time it may be easier to view the issues of social justice raised by these changes, but we of course have to act in the present.

Social justice and the practical application of its principles and values (autonomy, dignity, equity in the allocations of resources and opportunities to life and to realise benefits in life) requires an ecology of diverse tools, actors and infrastructures that can enable such collaborative efforts. Many of the problems we care about are complex, wicked. We believe advances in AI can have a role in helping shape positive responses to these problems. Yet there are significant challenges as to the way AI can contribute to more just forms of communal living. Challenges as to the creation of the conditions that may enable the many (rather than the few) to realise benefits from this technology.

The confluence of AI and social justice

AI is already having a profound impact on social issues across the world.



As has already been alluded to in issues of racial bias in AI systems. In particular, there has been extensive analysis of the COMPAS system used to assess the risk of re-offending (recidivism) in US courts when making parole; according to one analysis, it appears that the system systematically scores black defendants as more at risk than white defendants even when all other factors are taken into account [AL16, AL16b]. There are also counter-arguments that look forward to AI systems as part of legal practice as they are less likely to be swayed by prejudice than a human judge [Wa17].

In China mass facial recognition, powered by image analysis algorithms, uses cameras in public places. This has found wanted criminals, but is also suspected of being used to suppress opposition groups. Similar technology is being adopted by police forces across the world. In the UK a landmark case ruled that a trial of facial recognition by South Wales police was illegal, but only because it did not meet standards of proportionality and validity [Re20], not a blanket ban. Since then, the same police authority (and others across the country) is continuing to use the technology, but with better safeguards [SW22].

AI effects personal lives deeply. It is used extensively in financial risk assessment, including loan applications, which can have major impacts not just on whether loans are agreed, but also the interest rates charged, exacerbating tendencies for the poorer and less advantaged in society to pay higher interest rates [Lu16].

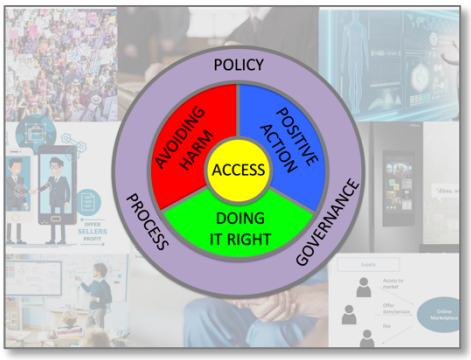
More profoundly, in war zones there is a fear that AI will be used not just to assist in life-or-death decisions (as is already happening), but to make those decisions autonomously. Drone swarms have already been used in Israel, although still under the ultimate control of a human operator [Ha21]. There is a considerable movement to ban the use of fully autonomous weapons, but countered by those arguing that their greater accuracy might prevent 'collateral damage'.

In 2009, as a response to the "*mouvement des gilets jaunes*" (yellow vest protests) in French, President Macron instigated a "Great National Debate", a series of face-to-face meetings across the country accompanied by a web-based systems to solicit widespread views. This exercise gathered more than 300,000 responses, far too many to deal with manually, so a company specialising in text analysis was recruited to extract key themes and issues [Fr19]. These were then used, as a form of focus group politics in the large, to shape French Government policy. One could say France was ruled by AI!

Making Sense of AI and Social Justice

We have found it useful to think about the interactions between algorithms and social justice under three main headings (see also diagram):

- ***avoiding harm*** – This is often the easiest to see, the things that are going wrong! More generally we need to identify existing or potential harms (such as issues of bias) and then seek ways to prevent, ameliorate or correct these.
- ***doing it right*** – This is where an organisation (commercial, government, or third-sector) wishes to use AI as part of their operations and want to ensure, from the start, that they do it in ways that do not cause problems for social justice. For example, this is one of the places where the explainable AI may be useful to help prevent unintended bias.
- ***positive action*** – Here typically an activist or grass-roots group wants to use AI for some positive social good. This may be to discover, evidence or highlight abuses. However, it may also include some application that helps those disadvantaged, for example, enabling communication between gig-workers.



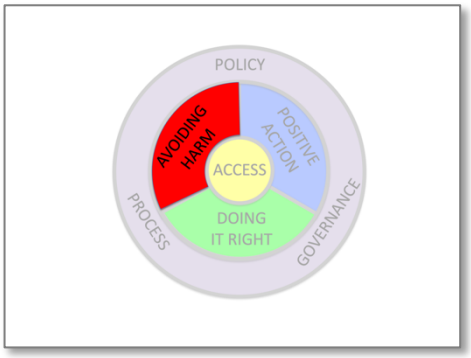
At the centre of this picture is access to digital technology, which undergirds so many other areas. This is partly about end-user access, where the existing social divides are often exacerbated by digital technology. To some extent these are generic digital problems. More AI-specific are issues facing organisations wishing to use AI for socially beneficial goals. The increasing computational costs (with associated environmental impact) can shut out all but the largest players from cutting-edge technology.

Around all this is context of government policy, legislation and regulation, internal governance within organisations and the processes by which AI is designed, deployed, and monitored.

We will look at the three main areas in turn as the inner and outer topics are more about governmental or top-down policy, so less immediately relevant for individuals working within AI development or social justice contexts to address. However, aspects will emerge as all the topics are interrelated.

Avoiding Harm

We'll start with avoiding harm as this is the most obvious of the three areas – the things that go wrong! This is not a new problem. In the 1960s Harvey Matusow began to catalogue "computer atrocities' [Ma68] and the potential for gender and ethnic bias in black-box machine learning was recognised in the early 1990s [Dx92].

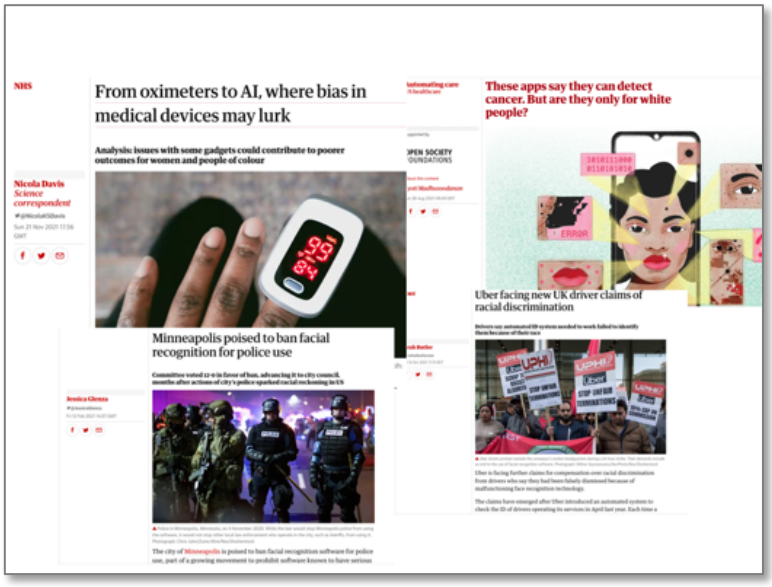


Sadly, today, examples of these problems and other breakdowns in AI deployment are continually reported in mainstream news.

One example, concerned oximeters used during Covid-19 pandemic, which may have contributed to higher deaths of non-Caucasian populations as the system was designed for "white people" and didn't appear to work as well on darker skin colours. Similar examples have cropped up elsewhere in healthcare with AI skin cancer detection apps working poorly on non-Caucasian-looking people [Da21].

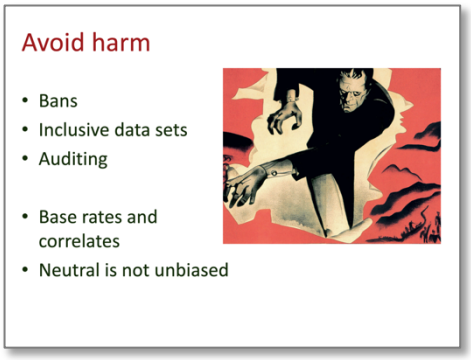
In general, medical devices and advice have often been developed based principally on information and studies drawing on white (and often male white) subjects and consequently fail for darker complexions. However, the problems go beyond the skin: different social and ethnic groups have different prevalence of particular kinds of disease or may present the same underlying condition in different ways.

Examples in other domains includes problematic uses of face recognition technology in policing and in public spaces, and more recently in sharing economy applications whereby uber drivers have been denied access to the platform (and therefore were left unable to work) because the face recognition technology didn't work as well given the colour of their skin [Ba21].



There is no single or problem-free solution. At the most extreme we could decide to ban or heavily curtail certain kinds of AI technology that do not comply or have not been audited/tested sufficiently on wider and diverse populations. This is indeed what we have done with other technology that has been considered too harmful – nearly every country has traffic speed limits and other safety legislation that trades off the most efficient use of the car (fast travel), with the social cost of dangers to the driver and other road users.

Happily, there is also a growing body of research literature and practical tools, such as IBM's AI Fairness 360 [AI22], which can help deal with issues of bias in AI. This can include methods to 'debias' data sets where there is suspected historical bias. One can also use sophisticated tools, or simple numerical auditing to detect bias, for example, wage-gap data for companies, or statistical analysis of court judgements.



Data sets certainly embody past human prejudice. Way before AI was being used to predict recidivism, analysis of parole decisions was shown to exhibit substantial racial bias [CM76]. Machine learning based on these past decisions would simply inherit the same prejudice.

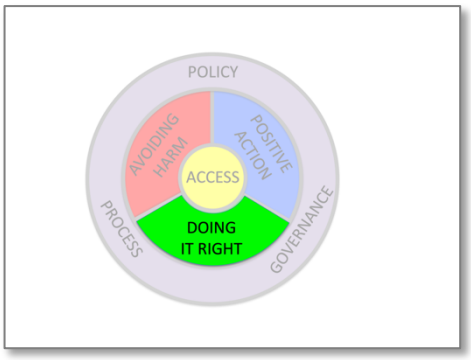
The answer to these issues has primarily been to make training data more inclusive, developing better and more effective auditing and it can be easy to assume that if one removes human bias from datasets, the eventual system will be discrimination free (e.g. [KC09,VE21]). However, this is often not the case, due to different *base rates*. For example, it is the case that, in the UK, pupils from lower socio-economic groups perform less well in school-leaving exams – social background is a predictor of academic success. If a machine learning algorithm were developed based solely on actual past performance (not human assessments of potential), it might easily use socio-economic background *as a predictor*. Of course, the reason for the differences in performance are themselves due to social conditions which embody past social injustice, but still this means that a system that tries to be neutral and based solely on actual performance may still end up biased.

The most egregious cases of this are avoided by removing potentially discriminatory variables, 'protected characteristics', such as race or gender from the data [Dx92,PR08,ZW13]. However, often these will correlate with other features of the data, and so discriminatory outcomes can still arise from apparently gender- or race-blind datasets. It is not enough to be neutral, one has to actively work to be fair. Information deriving from protected characteristics can be critical to make fair decisions that consider how structural conditions and features of the world play a role in anyone's effective opportunities to realise benefits from any resource available (equity). In a recent article, Hoffman shows that antidiscrimination discourse in data and AI can obscure the conditions that reproduce and normalise disadvantage [Ho19]. In order words, to actively work for social justice, one has also to question assumptions as to the factors that constitute advantage or disadvantage in the first place.

These issues were also highlighted in a recent article by Babu and Shahin [BA21] who charted the evolution of discourse in the proceedings surrounding Facial Recognition's ban in California in connection to the Body Cam Accountability Act. They charted how the discourse at first included the role of historical societal biases in the reproduction of discrimination through Facial Recognition technology; but then ended up framing the issues as a mere function of machine learning inaccuracies of the technology itself. The implication of this, they argued, has been a shorter ban of three years for Facial Recognition technology, as the technology "would, after all, improve – as technology is always expected to – and inaccuracies would reduce and eventually go away" [BA21 p.237]. Unfortunately, the framing of issues as narrowly technical ended up foreclosing spaces for much needed public debates and understandings on structural and historical racial bias and how these feed in AI deployments. It also discouraged discussions on Facial Recognition technologies' relationship to civil liberties, and the costs and trade-offs that adopting this technology imply.

Doing It Right

Many organisations do want to be fair and embody principles of social justice in their activities. Sometimes, this is part of the fundamental ethical position of the organisation, sometimes because the organisation just wants to avoid bad press.

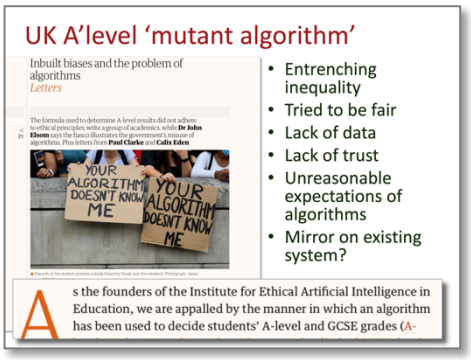


As is evident from the preceding discussion, this is not a simple task, and those who try may still end up castigated! As an illustration we'll look at a case study that highlights several common issues.

Case Study: UK's school exam 'mutant algorithm'

In Spring 2020, due to the Covid 19 pandemic, UK schools were closed to all except the children of key workers. It was decided early on that this would mean the normal A' level exams, administered at school leaving age (usually 18), would have to be cancelled. However, these exam results are used both for future jobs and university entrance. OFQUAL, the body responsible for overseeing UK exams, was tasked with coming up with an alternative way to deliver grades, based on teachers' predicted grades for their pupils, which are collected routinely every year to help universities make initial (usually conditional offers). However, it was known that predicted grades might overstate actual performance and so might not be fair both within the year (if different schools or teachers were more or less generous in their grading), or compared with previous years (if predicted grades were over-generous on average).

Ofqual created an algorithm that combined various factors including the predicted grade and statistics on the performance of each school in previous years. This was then used to create a corrected grade that could be higher or lower than the teachers' predictions.



When these predicted exam results were published in the summer of 2020, it created uproar.

The system placed constraints on how many pupils could achieve certain grades and based its outputs on a schools' prior performance, downgrading around 40 per cent of predicted results (although still allocating overall results far higher than previous years). The data across centres/schools was not equal and different centres had applied different principles to its standardization.

Individual pupils felt that their grades had been unfairly reduced, sometimes substantially, and furthermore analysis of the overall effect showed it was operating unfairly between different parts of society. Overall, the algorithm allowed a level of 'grade inflation' compared with previous years,

so that there were more top grades overall. However, analysis by journalists found that the increase amongst those going to fee-paying schools was around twice as high as that for state-funded schools [DM20]; that is the algorithm seemed to *entrench inequality*.

There followed media furore and protests by pupils outside parliament. Government ministers blamed Ofqual and the Prime Minister described it as 'mutant algorithm' [Co20]. Following the uproar from students and advocacy groups who took Ofqual and it's system to court [Fx20] – the system was withdrawn and Ofqual ended up utilising the teachers' predicted grades. Ofqual later took the highly unusual step of publishing a 318 pages report, which provided explanations of the logics, goals and objective of the system [Of20].

The sad thing is that Ofqual were *trying to be fair*. As soon as they knew the exams were to be cancelled, they gathered data on potential bias in teachers' predicted grades (aka. centre-assessed grades) and produced a public report on their findings [LW20]. This included some data from the UK, but also from other countries. The remarkable thing reading this report is how *little data* they had to go on. One might have thought that the body responsible for oversight of qualifications would be constantly monitoring data about social, racial and gender equality, but in fact this was limited to a few targeted research projects. Possibly this is because the regular collection of protected characteristics would potentially be seen as problematic.

Another factor evident in the whole process is a mutual *lack of trust*. Government and Ofqual did not trust the teachers, the pupils did not trust Ofqual and the whole rigour of the exam system is in part built on distrust of the pupils themselves. Exam grades are effectively a replacement for personal relationships.

Looking back at the algorithm itself. One of the problems was the *unreasonable expectations* of the algorithm – it was being asked to do an intrinsically impossible task. Without the actual data on exam results, the belief was that it was possible to create an algorithm that would in a sense fairly (at an individual level) recreate the missing exam grades, whilst also being fair (in a group sense) within and between cohorts. The algorithm was expected to create data out of thin air. It later transpired that early in the process Ofqual had realised this and advised government to scrap the exams entirely rather than attempt to create an algorithmic corrections, but were overruled.

However, perhaps the most enlightening aspect of the story is the way in which the failings of the algorithm cast a *mirror on the existing systems*.

The Guardian article reporting in August 2020 highlighted that the growth in top grades in independent (that is fee-paying) schools was 4.7% compared with 2% for state schools. However, this was an uplift to the previous year where around 45% of pupils in independent schools saw growth in top grades compared to around 20% of those at state schools, that is both had an uplift of 10% of their previous levels. It wasn't so much that the algorithm was acting unfairly but that the existing A level system is deeply socially divisive.

Finally, one of the reasons why this caused so much discord is that A level results are critically important to pupils. For those continuing to university the grades determine which university you can attend or whether you are admitted at all. For those going into the workplace, these are the grades that will be on their job applications for years to come. These grades quite literally shape the rest of the students' lives. Even in a normal year, a cold at the wrong time of year or a difficult time at home can make a difference to a student's life that will last years. This can be a personal tragedy for anyone, but also it is evident both from the Ofqual report and the raw grade figures that these impacts to not fall uniformly in society.

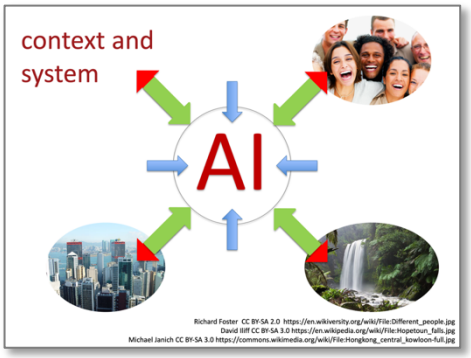
As a postscript to this story, after the dust settled Ofqual updated the literature review from 2020. This revised report, based on UK and overseas evidence, concluded that centre-assessed grades were likely to disadvantage both poorer pupils and those with special educational needs [LN21]. Researchers at UCL also did a post-hic survey which highlighted the disproportionate impact of Covid

on socially disadvantaged young people [AM21]. This survey included assessing the impact of the last-minute change from the Ofqual algorithm to using teachers' grades. It found that dropping the algorithm had led to 15% higher uplift for pupils with graduate parents [AM21b] – that is dropping the algorithm had yet further increased social disparity.

In other words, despite it being a bad idea in the first place, despite the public protests, government approbation and media vilification, Ofqual probably did a good job after all.

Context and system

Compared with AI machine learning systems, the A Level algorithm was quite simple and ultimately easy to understand [He20]. However, it has highlighted key points that are if anything more important when dealing with deep neural networks or other algorithms where we may need complex explainable AI techniques even to begin to interpret the outputs.



Crucially the focus for this story falls on the algorithm itself, whereas of course it was the wider impacts of the algorithm which mattered. If A Levels were less important, the accuracy of the algorithm would matter less. As with any socio-technical system, it is the overall outcome for the system as a whole that matters, not the specific outputs of the computer component.

Every system we design first into a rich context, this includes:

- ***social*** – the network of human relationships including trust, emotional effects and long-term human impacts;
- ***spatial and organisational*** – where and how the system is deployed including how the outputs will be interpreted, how they will be used;
- ***environmental*** – not least the growing understanding of the carbon cost of excessive computation.

Effective AI design for social justice is not just about avoiding bias in machine learning (although that is important), but understanding how the eventual system will fit into this larger matrix both within an organisation and beyond.

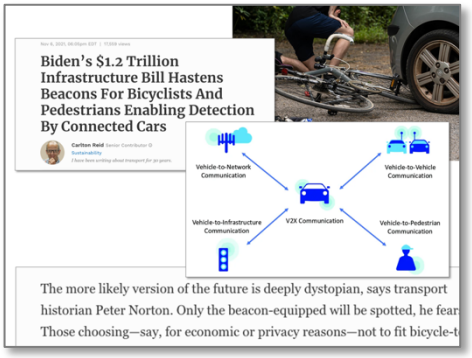
As we have seen, this is often difficult, partly because attempts to reduce unfairness in one aspect may lead to other forms of unfairness [CA21] (as was evident in the case study) or other unforeseen consequences.

More broadly, AI systems and their components (software, data, hardware) are shaped by wider socio-economic, political systems, and organisational cultures, which determine the extent to which people can realise benefits from this technology or any service that are being transformed with AI.

Existing ways of doing things, seeing the world, distributing resources and constraints, shape the logic of AI systems, and the generation of data, software, and hardware, and their goals, which —if left unchallenged, may reproduce assumptions and inequality.

Significant in all of this is the question of the kind of algorithms that we might choose to create, and for which purpose. The choices that we make as to the kind of problems that we respond to with AI are in and of themselves ethical matters. For example, AI might be usefully employed to help expose current flaws and structural inequalities in the UK schooling system, rather than trying to optimise already unfair systems and fraught processes. In this sense, we might choose to design AI to better understand current inequities and guide action as to how we can change a broader social system to advance social justice, rather than to advance existing systems that work for a privileged subset of society. We explore this in the next case about driverless cars.

Case study: Biden's bicycle beacons



In 2021 as part of President Biden's large scale infrastructure bill, was a proposal to modernise road infrastructure by providing vehicle to everything (V2X) technology. This would involve equipping bicycles (and potentially pedestrians) with transponder beacons that can be spotted automatically by sensor-equipped cars [RE21].

The case of this transformation appears to be that the technology will dramatically improve safety and reduce deaths on the road -- cars, and in particular autonomous cars, could detect vulnerable road users and avoid accidents. Yet, it also raises questions as to whose life is to be safeguarded in this scenario.

Rightly so, people have already pointed out how exclusionary realising such vision would be. So that for this to work – everyone would need a smart phone, or beacons (pedestrian or cars alike). It would be those who can't afford the technology whose life will be at greater risk (as highlighted in the article [RE21]). Bringing such vision to life, would also shift risk and responsibility for road accidents onto those who do not wear such beacons (by voluntary or enforced choice due to socio-economics). Furthermore, this then makes the possession of such a beacon virtually mandatory with all the attendant risk of pervasive surveillance.

More broadly (and paradoxically perhaps) the proposals appear to be centred on designing infrastructures that work for AI and the privilege of owning a driverless car, a mobile phone or beacon technology—rather than people. Where, in this vision, those who will suffer the most are those who are already economically marginalised and excluded in our societies. It also begs the question as to the necessity of such innovations and their ecological implications, in a time where innovating towards net-zero should be of critical importance. Who or what might benefit from AI innovations? Whose interests are being served?

Unexpected consequences

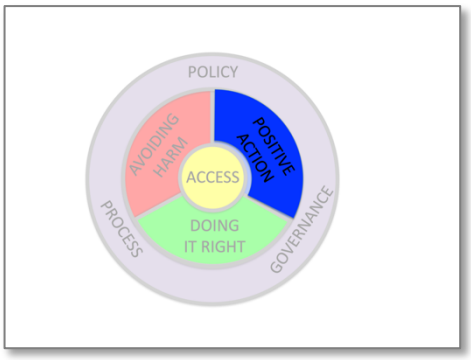
In the Fintech sector all players recognise the importance of ensuring fair algorithms, and some are actively working to ensure that they meet standards well beyond current legal and regulatory standards. Starling Bank's ethics statement includes a commitment to avoid bias and to ensure "*the benefits of our technology are able to be shared equitably across all parts of society*" [St22].

These are very positive signs, but of course the success of online banking, accelerated by health concerns due to Covid-19, have led to an overall reduction in the use of cash and of face-to-face banking services. In the UK, the number of bank branches halved between 1986 and 2014 and reduced 30% further since [UK22]. This has a major impact on those who find it harder to access or use online services, particularly older and other already disadvantaged groups. Furthermore, the closure of bank branches means there are less ATM machines, particularly in poorer areas of the country. Where ATMs are available, rather than the free service provided directly by banks, they are third-party machines in small shops on a pay-for-use basis. A typical charge is £1.50 for a cash withdrawal, independent of amount. For a well-off person withdrawing £100 or £200 at a time, this is relatively insignificant, but if you are less wealthy and withdrawing perhaps £10 or £20 at a time this is a major additional cost on an already tight budget.

Of course, for each bank this is hard to manage as they are quite reasonably trying to offer the best service to their customers. However, the nature of social justice often requires collective action whether by companies in a sector agreeing a voluntary scheme (e.g. a small levy on online transactions to subsidise ATM networks) or some form of regulatory or government response.

Positive Action

Finally, we come to positive action, where some form of community or activist group uses AI technology to create a positive social good or challenge some negative aspect. There are numerous broad initiatives in the area, including an annual UN sponsored Summit (https://www.itu.int/en/ITU-T/AI/) and non-profit/social enterprises (e.g. https://www.aiforgood.co.uk/, https://ai4good.org/).



In addition, there are numerous specific projects addressing particular needs, for example, the OSR4Rights project (https://osr4rights.org), which is using a variety of methods, including facial recognition, to help organisations such as the UN to gather evidence for war crimes and other human rights abuses [MM22].

Case study: HURIDOCS

Our case study here comes from the way ML tools have been used to help the work of advocacy groups – such as the Human Rights Information and Documentation Systems (HURIDOCS). HURIDOCS is a Geneva-based NGO that helps human rights lawyers and organisations manage and analyse data and collections of evidence they can use for social actions that support their causes. One of the ways HURIDOCS has done this work is to develop an open-source application—Uwazi, that enable organisations build and curate databases of human rights information (https://uwazi.io/). One particular challenge that human rights organisations face, beside building such databases, is also updating, curating and maintaining them as to provide information access in a timely and efficient manner.

In 2018, HURIDOCS won a 1M US dollar grant from the Google Artificial Intelligence Impact Challenge [Go18], which included participation onto the Google fellow programme, enabling Google employees to do full-time pro bono work with them for six months. The grant enabled HURIDOCS to develop and implement Machine Learning features on their existing open-source application Uwazi (https://uwazi.io). The ML features would help human rights organisation import automatically new human rights information when made available as well as curate this information database through extraction of paragraphs and providing suggestion for appropriate labels to access and curate human rights information they could use for their causes within weeks, rather than months. As part of this grant HURIDOCS and google fellows worked to improve access to information from the Universal Periodical Review (UPR). The UPR at the UN Human Right Council is a process that involve each member state declaring what actions they have taken to improve the human rights situations

in their countries and to fulfil their human rights obligations. Therefore, the UPR consists of long documents containing recommendations for improvements on the UN website. In order to facilitate access to this information to advocates, researchers and diplomats, the UPR Info database was created providing recommendations, with taxonomies like Human Right issues, recommending state and receiving state. Still the curation of the database is time-intensive and requires supervision and training. So, they used participatory methods to develop a human-in-the loop system that employed machine learning to extract paragraphs and make suggestions for labels.

Ultimately the system didn't solve human rights issues but played a role in the ecology of tools that human rights advocate can utilise for their causes.

Reflection – relational AI

Unpacking implications from examples

These examples, speak of various significant issues with AI:

- *Bias and standards*. Our expectations that with AI we can get rid of biases and that through standardization it is possible to achieve de facto equality (even where data is altogether missing). The cases showed also assumptions around representation in datasets, but also standardization across datasets, where these are assumed to be produced in equal ways (which is often not the case). In other cases, assumptions around neutrality of data leading to unbiased outputs, can obscure questions as to the structures and wider systems that produce advantage or disadvantage. In the A level case, ultimately the students wanted to be judged according to the particularities of their contextual learning journeys and their capabilities—that is equity not equality.
- *Trust & responsibility*. Perhaps more subtle is also how these examples, throw up important reflections concerning trust and responsibility. The AI level controversy is exemplary of what can be defined as a sort of spiral of mistrust generating more mistrust. Ofqual didn't trust the teachers to predict grades based on their knowledge, but placed trust in the AI system, whose contested outputs generated mistrust in the AI system. Conversely, Biden's vision for bikes' beacon as a way to achieve road safety shifted responsibility for road safety onto the individual. In both cases AI systems redistributed and refigured agency and responsibilities. These examples draw attention to the ways AI systems can be configured to reinforce existing relationship of trust or mistrust and the enacting of responsibility as an individualised matter, rather than a collective affair.
- *Goals and benefits.* The cases showed how an excessive focus on making the AI work, can side-line the importance of considering contexts and people, leading to systems that attempt to optimise around goals, which disadvantage those who are meant to benefits form these innovations. As such, some of these cases are also powerful reminder of the inequities at play in the way AI systems' goals and risks associated are defined, and by whom, and how definition of goals and risks lead to particular ways to model a technological solution. In this sense it becomes important to include people who use or are at the receiving end of these innovations, in the articulation of goals, norms and values that should guide the decision any AI system is supporting.
- *Public understanding of AI issues and broader transformations*. Another important aspect emerging from these cases concern the way the framing of issues arising from AI deployments in societies can open or foreclose constructive public understanding and discourse on AI. We have seen how a focus on technical accuracy can block out broader questions and discussions around

the broader systems the AI is built upon or discussions around surveillance and AI's place in societies. The AI level case is compelling a one of the fewer examples of expert's extensive reporting of a system gone wrong and students' population mobilising and reporting how they felt through mainstream and online media channels. The example shows gaps between experts and people's discourse and understanding on these issues. Closing this gap and findings new ways to enable constructive dialogue would be important to help create and sustain a critical mass of experts and publics that can shape how AI systems should be governed.

Breakdowns as opportunities for alternatives to emerge: relational thinking in AI

Taken together, the examples we've seen of AI breakdowns are important not only because they make visible and palpable the actual and potential consequences of careless/irresponsible design and deployment (e.g. detached from the historical and cultural contexts in which data and software are generated); but also because they expose the knowledge base, worldviews, and values that characterise our information infrastructures, which are often taken for granted and that have come to be normalised. They expose the underlining (already) unequal societal structures that uphold, feed and shape AI systems. They expose how in turn, AI systems shape, feed and uphold our unequal and unfair societal structures. For example, the way AI plays a role in the distribution of opportunities and benefits in life, which critically are also opportunities to life and survival.

AI is conceived and deployed in a way that at best leaves existing power relationships in society unchanged, and at worst exacerbate their unequal dynamics—whereby those with little power, have even less. People and their proxies (data) are framed as something to be managed and decisions as something that are done for or to them with little or no routes for accountability.

While inclusive datasets, bans, etc. are useful ways of avoiding negative externalities and harm, this may be only a partial and limited way of responding and considering the wider issues at stake. More troubling perhaps, they can leave the root causes and power dynamics at play in today's unequal realities unchanged and unquestioned – thus limiting opportunities for conversations and imaginations on what the role of AI advances for social justice could or should be, instead. To think of AI relationally is also to ask ourselves: how might AI be designed and configured in ways that foster more constructive (rather than oppressive) relations among people, service providers, and governments, and that redress the balance in unequal power relations. To this end, conversations and visions of AI's place in societies should be centred around the interests and voices of those at the receiving end of AI innovations to help re-articulate how this particular technology can support values and goals for social justice, and how.

AI for Social Justice – challenging but important

As we've seen, it is not easy to design AI that avoids creating social harms, let alone promoting social good, but it is of course crucially important. It is important for society and individual people who are already affected day-to-day, for whom, if left to proceed without care, AI is likely to disadvantage the weakest. It is also important for the technology sector, if this is not managed adequately the regulatory responses are likely to be blunt instruments that restrict the positive potential of AI as well as curb excesses.

For more on these issues keep an eye on our website and watch out for our new book that we are hoping to complete soon.

References

[Ad21] Richard Adams (2021). Ofqual wanted to scrap last year's A-levels, says former chair. The Guardian, 14 Jun 2021. https://www.theguardian.com/education/2021/jun/14/ofqual-wanted-to-scrap-last-years-a-levels-says-former-chair

[AI22] AI Fairness 360. (Open source toolkit) Accessed 12/8/2022. https://aif360.mybluemix.net

[AM21] Anders, J., Macmillan, L., Sturgis, P. and Wyness, G. (2021). Inequalities in young peoples' educational experiences and wellbeing during the Covid-19 pandemic (CEPEO Working Paper No. 21-08). Centre for Education Policy and Equalising Opportunities, UCL. https://EconPapers.repec.org/RePEc:ucl:cepeow:21-08.

[AM21b] Anders, J., Macmillan, L., Sturgis, P. and Wyness, G. (2021). The 'graduate parent' advantage in teacher assessed grades. (blog) UCL, Centre for Education Policy and Equalising Opportunities. 8 June 2021. https://blogs.ucl.ac.uk/cepeo/2021/06/08/thegraduate-parentadvantageinteacherassessedgrades/

[AL16] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In Ethics of Data and Analytics (pp. 254-264). Auerbach Publications.

[AL16b] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bia. ProPublica. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[Ba21] Sarah Butler. 2021. Uber facing new UK driver claims of racial discrimination. The Guardian, 6 October 2021. https://www.theguardian.com/technology/2021/oct/06/uber-facing-new-uk-driver-claims-of-racial-discrimination

[CM76] Leo Carroll and Margaret E. Mondrick (1976). Racial Bias in the Decision to Grant Parole. Law & Society Review. Vol. 11, No. 1 (Autumn, 1976), pp. 93-107

[CA21] Cooper, A. F., Abrams, E., & Na, N. (2021). Emergent unfairness in algorithmic fairness-accuracy trade-off research. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 46-54).

[Co20] Sean Coughlan (2020). A-levels and GCSEs: Boris Johnson blames 'mutant algorithm' for exam fiasco. BBC News. 26 August 2020. https://www.bbc.co.uk/news/education-53923279

[Da21] Nicola Davies. 2021. From oximeters to AI, where bias in medical devices may lurk. The Guardian, 21 November 2021. https://www.theguardian.com/society/2021/nov/21/from-oximeters-to-ai-where-bias-in-medical-devices-may-lurk

[Dx92] A. Dix (1992). Human issues in the use of pattern recognition techniques. In Neural Networks and Pattern Recognition in Human Computer Interaction Eds. R. Beale and J. Finlay. Ellis Horwood. 429-451.

[DM20] Pamela Duncan, Niamh McIntyre, Rhi Storer and Cath Levett (2020). Who won and who lost: when A-levels meet the algorithm. The Guardian, 13 Aug 2020. https://www.theguardian.com/education/2020/aug/13/who-won-and-who-lost-when-a-levels-meet-the-algorithm

[EC21] Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts. European Commission, COM(2021) 206 final, 2021/0106(COD), Brussels, 21.4.2021

[Fx20] Foxglove Legal (2020). We put a stop to the A Level grading algorithm! Foxglove Blog, 17th August 2020. https://www.foxglove.org.uk/2020/08/17/we-put-a-stop-to-the-a-level-grading-algorithm/

[Fr19] France 24 (2019). What will France do with 'National Debate' data? France 24. 03/03/2019. https://www.france24.com/en/20190302-france-great-national-debate-data-artificial-intelligence-politics-yellow-vests

[Go18] Google.org (2018). Google AI Impact Challenge, 2018: Working together to apply AI for social good. Accessed 13/8/2022. https://impactchallenge.withgoogle.com/ai2018

[HN09] A. Halevy, P. Norvig and F. Pereira, "The Unreasonable Effectiveness of Data," in IEEE Intelligent Systems, vol. 24, no. 2, pp. 8-12, March-April 2009, doi: 10.1109/MIS.2009.36.

[Ha21] David Hambling Israel used world's first AI-guided combat drone swarm in Gaza attacks (2021). New Scientist, 30 June 2021. https://www.newscientist.com/article/2282656-israel-used-worlds-first-ai-guided-combat-drone-swarm-in-gaza-attacks/

[He20] Alex Hern (2020). Ofqual's A-level algorithm: why did it fail to make the grade? The Guardian. 21 Aug 2020. https://www.theguardian.com/education/2020/aug/21/ofqual-exams-algorithm-why-did-it-fail-make-grade-a-levels

[Ho19] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse, Information, Communication & Society, 22:7, 900-915, DOI: 10.1080/1369118X.2019.1573912

[KC09] F. Kamiran and T. Calders (2009). Classifying without discriminating. 2nd International Conference on Computer, Control and Communication, 2009, pp. 1-6, DOI: 10.1109/IC4.2009.4909197.

[Lu16] Sarah Ludwig (2015). Credit scores in America perpetuate racial injustice. Here's how. The Guardian, 13 October 2015. https://www.theguardian.com/commentisfree/2015/oct/13/your-credit-score-is-racist-heres-why

[LW20] Ming Wei Lee and Merlin Walter (2020). Equality impact assessment: literature review. Office of Qualifications and Examinations Regulation (Ofqual) April 2020.

[LN21] Ming Wei Lee and Paul Newton (2021). Systematic divergence between teacher and test-based assessment. Office of Qualifications and Examinations Regulation (Ofqual), 17 May 2021. Ref. Ofqual/21/6781. https://www.gov.uk/government/publications/systematic-divergence-between-teacher-and-test-based-assessment

[Of20] Ofqual (2020). Awarding GCSE, AS & A levels in summer 2020: interim report. Office of Qualifications and Examinations Regulation (Ofqual)13 August 2020. Ref: Ofqual/20/6656/1. https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report

[Ma68] Harvey Matusow (1968). The Beast of Business: a Record of Computer Atrocities, Wolfe, ISBN: 0723400601

[MM22] Murray, D., McDermott Rees, Y., & Koenig, K. (2022). Mapping the Use of Open Source Research in UN Human Rights Investigations. Journal of Human Rights Practice, https://doi.org/10.1093/jhuman/huab059

[PR08] Pedreshi, D., Ruggieri, S., and Turini, F. Discrimination- aware data mining. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discov- ery and data mining, pp. 560–568. ACM, 2008.

[Ra71] Rawls, John (1971). A Theory of Justice. Harvard University Press

[Re20] Jenny Rees (2020). Facial recognition use by South Wales Police ruled unlawful. BBC Wales, 11 August 2020. https://www.bbc.co.uk/news/uk-wales-53734716

[Re21] Carlton Reid (2021). Biden's $1.2 Trillion Infrastructure Bill Hastens Beacons For Bicyclists And Pedestrians Enabling Detection By Connected Cars. Forbes, Nov 6, 2021. https://www.forbes.com/sites/carltonreid/2021/11/06/bidens-12-trillion-infrastructure-bill-hastens-beacon-wearing-for-bicyclists-and-pedestrians-to-enable-detection-by-connected-cars/?sh=5e9fce735a3d

[SW22] South Wales Police / Heddlu De Cymru (2022). FAQs about Live Facial Recognition Technology., Accessed 12/8/2022. https://www.south-wales.police.uk/police-forces/south-wales-police/areas/about-us/about-us/facial-recognition-technology/faqs-about-live-facial-recognition-technology/

[St22] Starling Bank (2022). Our ethics statement. Accessed 13/8/2022. https://www.starlingbank.com/about/ethics-statement/

[UK22] UK Parliament (2022). Statistics on access to cash, bank branches and ATMs. Research Briefing. House of Commons Library. 25 July 2022. https://commonslibrary.parliament.uk/research-briefings/cbp-8570/

[VE21] Verma, S., Ernst, M., & Just, R. (2021). Removing biased data to improve fairness and accuracy. arXiv preprint arXiv:2102.03054. https://arxiv.org/abs/2102.03054

[Wa17] Caleb Watney (2017). It's time for our justice system to embrace artificial intelligence. Brookings, July 20, 2017. https://www.brookings.edu/blog/techtank/2017/07/20/its-time-for-our-justice-system-to-embrace-artificial-intelligence/

[ZW13] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In International Conference on Machine Learning, pp. 325–333, 2013.