Sufficient Reason

algorithmic bias and transparency

Alan Dix

Computational Foundry
Swansea University

http://alandix.com/academic/talks/sufficient-reason-2018/

These are notes of a keynote originally given at the Workshop on HCD for Intelligent Environments, BHCI, Belfast, 3rd July 2018 [Dx18].

Abstract

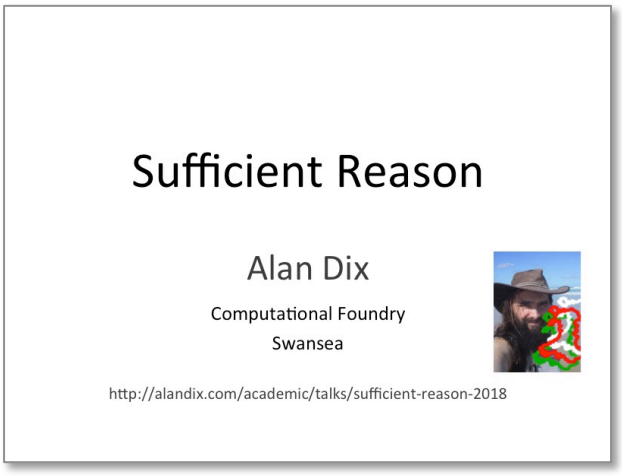
A job candidate has been pre-selected for shortlist by a neural net; an autonomous car has suddenly changed lanes almost causing an accident; the intelligent fridge has ordered an extra pint of milk. From the life changing or life threatening to day-to-day living, decisions are made by computer systems on our behalf. If something goes wrong, or even when the decision appears correct, we may need to ask the question, "why?"

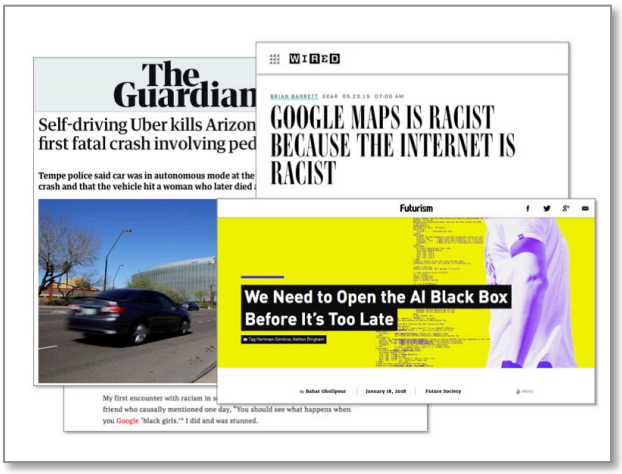
In the case of failures we need to know whether it is the result of a bug in the software, a need for more data, sensors or training; or simply one of those things: a decision correct in the context, that happened to turn out badly. Even if the decision appears acceptable, we may wish to understand it for our own curiosity, peace of mind, or for legal compliance.

In this talk I will pick up threads of research dating back to early work in the 1990s on gender and ethnic bias in black-box machine-learning systems, as well as more recent developments such as deep learning and concerns such as those that gave rise to the EPSRC human–like computing programme.

In particular I will present nascent work on an AIX Toolkit (AI explainability): a structured collection of techniques designed to help developers of intelligent systems create more comprehensible representations of the reasoning. Crucial to the AIX Toolkit is the understanding that human–human explanations are rarely utterly precise or reproducible, but they are sufficient to inspire confidence and trust in a collaborative endeavour.

Keywords: artificial intelligence, interaction design, bias, explainability, user experience, UX, AI, HCI, trust



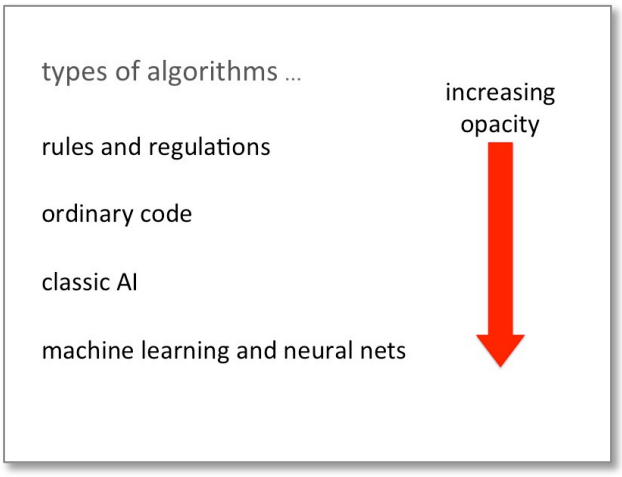


The last few years have seen a growing optimism in the power of artificial intelligence and machine learning to address problems that were once seen as essentially requiring human understanding. In language processing, Google uses the power of vast volumes of data to create translations that appear to understand the original text [HN09]; in vision, algorithms are able to pick out the partially obscured faces of suspects in vast crowds; in commerce, Amazon seems to be able to suggest books or videos with apparently uncanny insight; in games, Go, once seen as unassailably complex, has been conquered by AI and the AlphaGo engine is creating new opening strategies for chess; and, on the roads, autonomous vehicles are escaping the lab and are expected to become commonplace within a few years.

However, this tide of optimism has been countered by an increasing number of stories of failures of various forms, from autonomous car fatalities to 'racist' search results [No18, LW18, Ba15, Bo19, Gh18]. Some are also laying part of the blame for growing intolerance and extremism in society on the algorithms behind social media and search engines, which, in the interest of giving

consumers what they want, are creating sounding boxes where we each see only views, news and 'facts' that agree with our own preconceptions.

This has led to calls for accountability and transparency of algorithms. Notably the European general data protection regulation [CEU16] demands that when algorithms make decisions that affect individuals, for example credit scoring or job shortlisting, these need to be capable of explanation, albeit, as Crabtree argues, often misinterpreted in scope [Cr19].



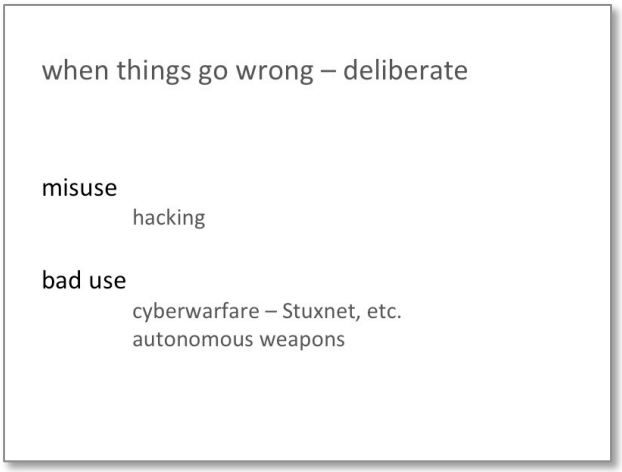
Although these issues are most stark when the algorithms involve machine learning, they arise with all sorts of algorithms, even the simplest. For example, in the 1980s a project intended to capture some of the legislation around welfare benefits found there were inconsistencies that had previously been overlooked [Le82, Le85].

In general, few understand the relatively small set of rules around taxation; for example, in the UK there are two forms of direct taxation, income tax (a general tax) and National Insurance (originally created to fund health services and pensions). The former is the 'headline' tax and governments often seek to minimise it; crucially it is progressive with a small amount of tax-exempt income, a standard-rate band of 20%, followed by 40% for higher income (over around £50,000), rising to 45% for the highest earners (over £150,000). What few people realise is that National Insurance drops from 12% to 2% at nearly the same point as the 40% band starts. That is, the main total tax rates are effectively 32%, 42% and 47%, far less progressive than most citizens realise, and this is before taking into account the regressive nature of many indirect taxes on goods, which tend to disproportionately target those with lower incomes.

As things get more complex, few programmers will attest to understand all the behaviour of their code. Indeed, advocates of formal methods in computing attempt to address precisely this issues, but these methods often prove too cumbersome for all but the most safety critical situations or toy problems in research contexts.

Classic symbolic AI is not so far from programming, although far more broad in terms of computational genres. Crucially many AI languages and notations are declarative; this can make them more clear in intent than ordinary code, but may also make the consequences of multiple interacting rules hard to predict. On the other hand AI techniques such as formal argumentation logics may offer ways to help other algorithms become more explainable.

Finally, most of the media coverage both positive and negative in recent years has concerned machine learning of various forms, from fairly simple frequency-based techniques on big data, to neural methods in deep learning. For these techniques the explanation as to *why* something has happened often comes down to "there was this shed load of data and this is what came out".



The press naturally focuses on the bad side of algorithms and machine learning. However, it is important to distinguish different forms of bad outcomes.

Some of these are clearly deliberate.

The first kind is deliberate misuse such as hacking. Here we may not blame the algorithms per se, but their weakness or vulnerability. In such cases we may seek better software design or security. For classic hacking the vulnerability is not in the complexity of intelligent algorithm itself, but the surrounding operating system, device drivers, etc. Once the hacker is 'in' they may subvert the software – modifying or replacing the code.

However, big data has led to more complex forms of vulnerability based on subverting the *data*. In the case of the Cambridge Analytica scandal, this was principally about using data in ways it was not supposed to be used; this is more strongly connected with privacy and personal control of data. However, the other aspect of this scandal was the way the resulting data was used to influence the US presidential elections. This was a fairly direct use of data, but often it can be less direct. Twitter Bots often deliberately create inflammatory Tweets on *both* sides of an issue; the aim is to increase re-tweets and hence the ranking of the channel, so that subsequent deliberately misleading or misinformative tweets (fake news) will have instant influence. A more citizen-led form of data

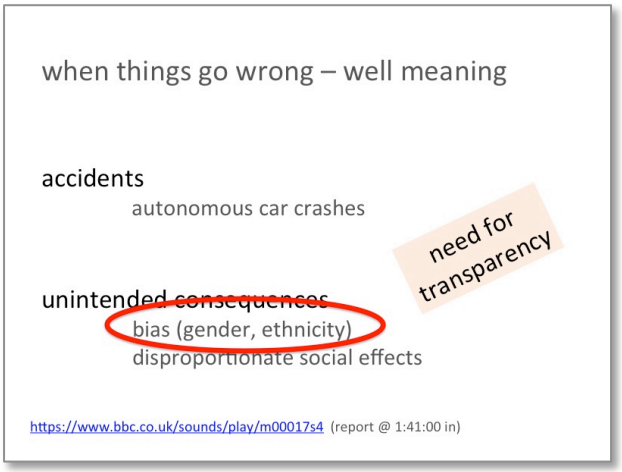
manipulation is used by campaigns to get everyone to do specific Google searches in order to make it have a particular auto-completion when you start to type a query such as ”Brexit”.

Some forms of deliberate ‘bad’ use are legal and may be regarded as acceptable depending on one’s ethical viewpoint.

Autonomous weapons have been widely condemned if there is not a human in the loop, and major scientists and industrialists have called for there to be an international ban akin to that on chemical weapons [HM15]. Oddly we accept bombs that explode at a fixed height, or guide themselves to a specific location using GPS; we also accept soldiers trained to obey orders without question. Perhaps unpacking what is so bad about autonomous weapons could help us unpack our attitudes to war itself?

Cyberweaponary has also attracted significant publicity. Some simply attack computer software or data, for example, Denial of Service (DoS) attacks, but others may be targeted to bring down infrastructure or even cause physical damage. The ethical attitude to cyberweapons does seem to be closely aligned to political acceptance of those wielding them.

A major recent concern has centred around Chinese-made routers and phones, and alleged Russian attempts at election interference are also often described as unacceptable cyberattacks. However, the first (publically known) case of cyberwarfare against physical infrastructure was Stuxnet designed by Israeli and US intelligence to degrade to Iranian nuclear programme by attacking centrifuges [Ku13]. Spread by USB memory sticks, Stuxnet showed that even Internet-isolated computers could be at risk and there are also rumours that it ‘escaped’ spreading to unintended targets with similar hardware including Ukrainian and Russian nuclear power stations [Vi13]. The US worries about Chinese routers may well be because a large proportion of US-made routers in the middle-east have been found to be infected with malware, which is assumed to be related to intelligence gathering on fundamentalist groups.



Sometimes things go wrong unintentionally, whether through ignorance, negligence or pure accident.

Autonomous car accidents have been widely reported. In some cases, particularly with Tesla vehicles, this has been because drivers have not understood the capabilities of cars that operate in semi-autonomous mode, but rely on the driver to maintain attention. Some suggest that this means that only fully autonomous vehicles should be allowed. However, there is a long history of partial autonomy from ABS to cruise control; so it may be that the answer is better design of the autonomous vehicle user interface, crucially ensuring that the 'driver' has a clear understanding of the momentary level of autonomy … and maybe also that the vehicle has a model of the driver's attention

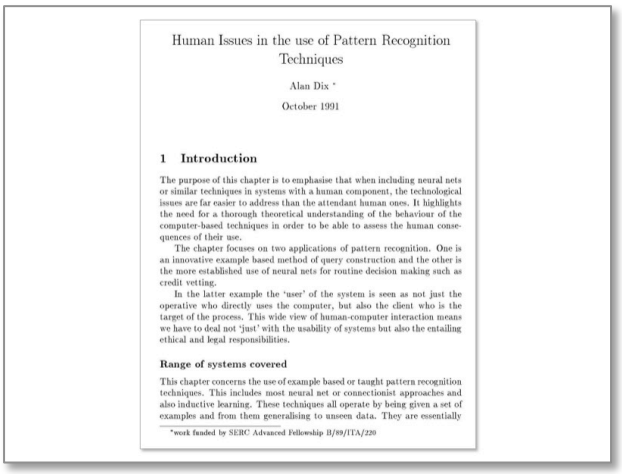
In other cases the vehicle has been in fully autonomous mode. Here the manufacturers and accident investigation authorities have to determine whether this was unavoidable (for example a person running into the road in front of the vehicle) or potentially preventable. In the latter case it is particularly important to be able to unpack the chain of sensing and decisions that led to the accident in order to see whether there are changes that could improve safety. The software for such vehicles is inevitably complex with interacting sets of rules and machine-learnt aspects, making such essential explanations difficult.

Even where there is no obvious cataclysmic 'accident', things can go wrong. The system appears to work and make suitable decisions, but are there unintended consequences?

One of these, and the main focus of the rest of this talk is the potential for unintended bias including gender or ethnic bias, in automated decision-making systems. There have been headline cases of this, notably Microsoft Twitter-bot Tay, that quickly learnt sexist, racist and anti-Semitic language [Gu16]; or cases of Google search returning gender stereotyped images, or auto-completion [La18]. However, potentially more worrying are the cases we don't notice: what are the factors that are being used to set your loan interest rate; or determine whether you are shown highly-paid job adverts [DT15, Ca15]?

Even where systems are utterly neutral, the impact of numerous design decisions may affect different groups disproportionately. For example, the UK's Universal Credit system is designed to unify and simplify welfare payments, but is completely computer based. Early trials showed that 49% of those eligible did not have internet access at home [CAB13]. In rural areas land-based broadband and mobile connectivity may be slow or non-existent, so that not just Universal Credit, but all forms of Internet-shopping, eGovernment, or other services are hard to access [MD14].

In 1991, in the beautiful surroundings of King's Manor in York, I attended a workshop on Neural Networks and Pattern Recognition in Human Computer Interaction. This was during one of the periods when there was growing confidence in AI and machine learning: early recommender systems were being produced, neural networks were showing promise in areas from image recognition to fusion reactor control [Gr91], and Allen Cypher's EAGER demonstrated that programming-by-example could be integrated into routine interactions [Cy91].



Human Issues in the use of Pattern Recognition Techniques

Alan Dix *

October 1991

1 Introduction

The purpose of this chapter is to emphasise that when including neural nets or similar techniques in systems with a human component, the technological issues are far easier to address than the attendant human ones. It highlights the need for a thorough theoretical understanding of the behaviour of the computer-based techniques in order to be able to assess the human consequences of their use.

The chapter focuses on two applications of pattern recognition. One is an innovative example based method of query construction and the other is the more established use of neural nets for routine decision making such as credit vetting.

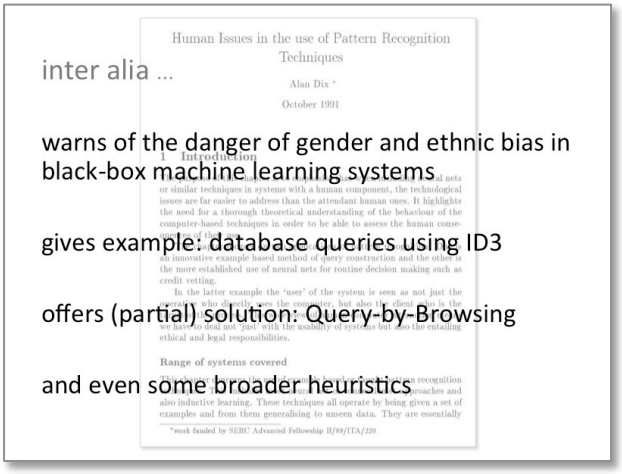
In the latter example the 'user' of the system is seen as not just the operative who directly uses the computer, but also the client who is the target of the process. This wide view of human-computer interaction means we have to deal not 'just' with the usability of systems but also the entailing ethical and legal responsibilities.

Range of systems covered

This chapter concerns the use of example based or taught pattern recognition techniques. This includes most neural net or connectionist approaches and also inductive learning. These techniques all operate by being given a set of examples and from them generalising to unseen data. They are essentially

*work funded by SERC Advanced Fellowship B/89/ITA/220

I gave the capstone talk for the day, looking beyond the immediate systems and techniques available at the time and towards future issues for the area. The talk subsequently was published as a chapter in a collection based on the day edited by Russell Beale and Janet Finlay [Dx92].

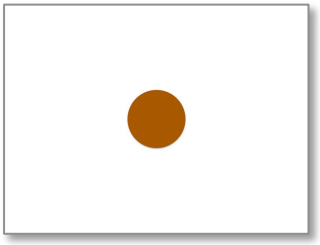


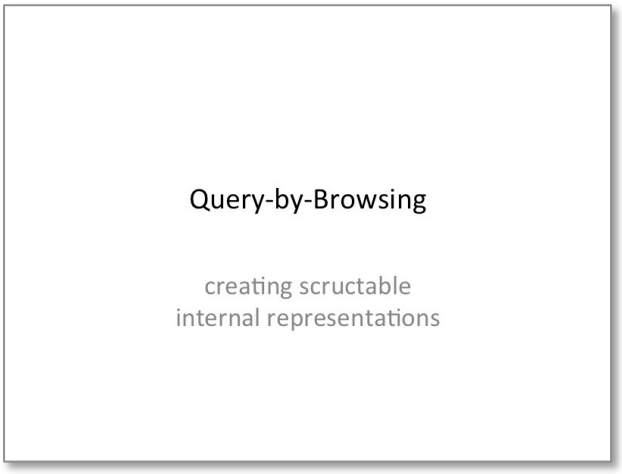
Amongst other things, the talk and chapter:

- warns of the danger of gender and ethnic bias in black-box machine learning systems
- uses an example of database queries generated using ID3 to demonstrate the potential problems
- offers a (partial) solution in the form of an envisioned system Query-by-Browsing , that enables a level of explanation
- outlines some broader heuristics for detecting, avoiding, or ameliorating bias

This now seems prescient. However, at the time, more than 25 years ago, this seemed not so much gazing into the far future, but highlighting problems that were likely to emerge in the next few years. Of course the heady days of the early 1990s AI bubble soon led to a period of disillusionment, sometimes called the 'AI Winter', that ran until the recent rise of big data, deep learning, and other technologies.

Now, 25 years on, these problems have come back with a vengeance.





The example system used through the 1992 paper was called Query-by-Browsing. In the paper it was a thought experiment to highlight potential problems and solutions, but was subsequently implemented [DP94] and is still available as a web demonstrator:

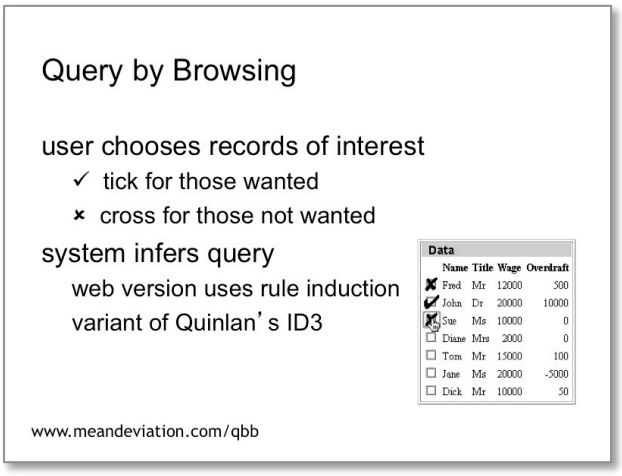
https://www.meandeviation.com/qbb/qbb.php

At the time, early news recommender systems were being developed using relevance feedback. A recommender for news has to be 'good enough', finding sufficient relevant articles to suggest and not showing too many irrelevant items. Precise accuracy is not required.

In contrast, when querying a database, say to select specific group of staff for a pay rise, it is usually important that the records selected are precisely those that are required. This is commonly achieved by writing an SQL query to select the required records, but this requires both technical expertise and the ability to frame one's requirements in precise logic. It would be nice to be able to use relevance feedback style interactions to select the records desired and then let that determine which staff receive the pay rise.

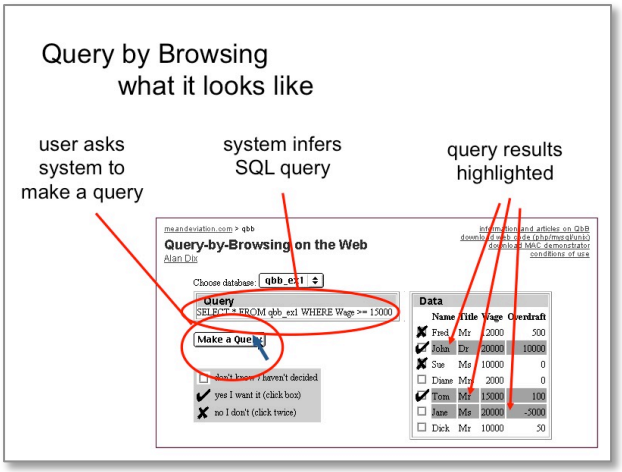
The technical challenge is to do this in a way that (a) you can be sure is updating precisely the right staff; (b) the rule used is one that does not violate any anti-discrimination legislation.

The record form of a database table is compatible with the input format of many machine-learning methods; however, most of these have relatively opaque learning algorithms and decision rules.



Query-by-Browsing attempts to address this, starting with relevance-feedback-style user-selection of records, but creating rules that are scrutable addressing requirements (a) and (b), and in the process highlighting the potential for biased results to arise that would be illegal in a less transparent system.

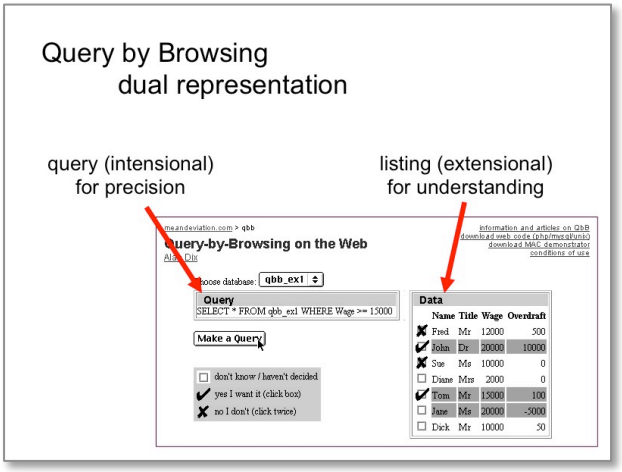
The original machine learning system chosen was a variant of ID3, Quinlan's original decision tree classifier [Qu86], extended to allow multi-column comparison criteria. This is also used on the web version; however, an intermediate version used genetic algorithms to create rules [Dx98].



Walking through the behaviour (web version shown):

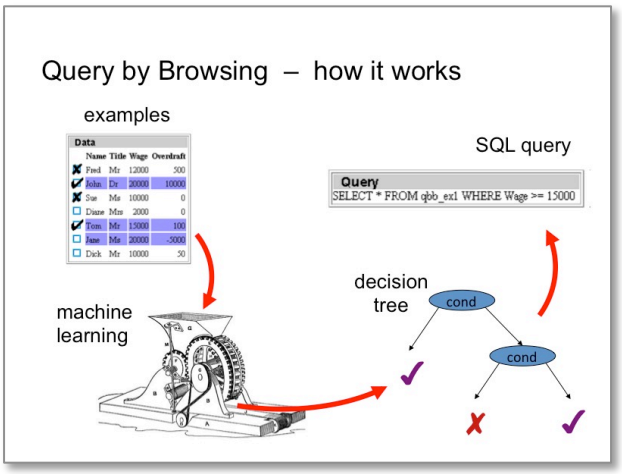
1. the user selects records of interest with a tick for those that are wanted and cross for those not required.
2. The user selects "Make a Query".
3. The system generates an SQL query that matches the desired records. (or RQBE in the case of GA version)
4. The query is displayed in the Query area and the records selected by the query are shown highlighted.

5. The user can select more examples and counter-examples to refine the query.



Note that the interface effectively includes two representations of the decision rule. In the Query area the decision tree is rendered as an SQL query giving an *intentional* representation; this is useful for precision, ensuring that conditions are exactly as required. In the List area the highlighted records form an *extensional* view of the rule, showing which records are chosen by it. This is particularly useful for complex and–or queries, or those including negation, which are known to be hard to interpret.

As well as allowing precision, the Query area makes the decision rule transparent. It is immediately obvious if the rules says, for example, 'SELECT * WHERE title="Mr"'. As we shall see later, this is not sufficient to ensure no bias, but certainly helps to uncover problematic decision rules.



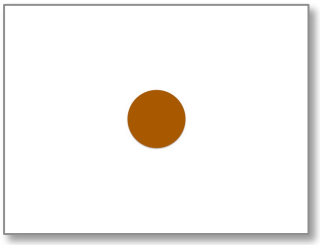
The diagram shows schematically the steps 'under the bonnet'. The examples chosen by the user are fed into a machine learning system that generates a decision tree or similar rules, and these are then rendered as SQL (or RQBE).

In the case of ID3 the top-down divide and nature of the machine learning algorithm is itself comprehensible, it is possible, albeit tedious, to go through the process by hand or read a trace of the system learning. In the case of genetic algorithms as a learning process the size of population and number of generations make the process opaque; however, the rules generated are still understandable.

This is crucial.

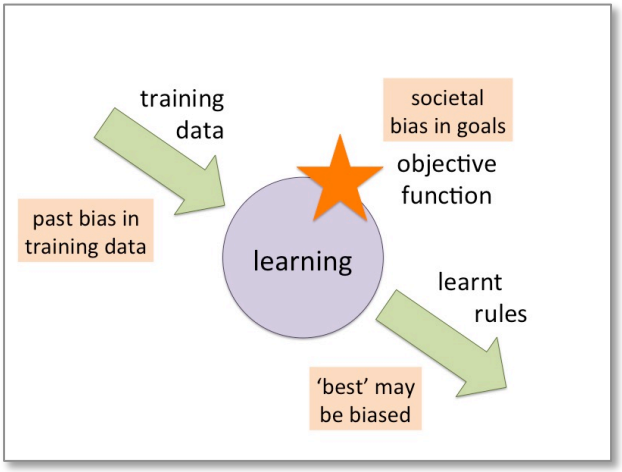
Think of a mathematician; the process of finding a proof may require trial and error, sparks of insight, generating intermediate lemmas. To fully describe and justify each step of this would be impossible. However, mathematicians do not attempt to represent *how* they came to a solution, but instead present a proof, a form of rational reconstruction of the actual mathematical process.

Similarly, it is often acceptable to have an opaque learning process so long as the rules generated are comprehensible.



We will now move on to consider different sources of bias. Crucially the 'bias' that we refer to in ethical or legal discussions is not the 'bias' used in technical discussions. In statistics a 'biased' procedure is one that in the long run returns a result that is not accurate, perhaps over- or under-estimating the true value. However, we shall see that even an algorithm that is entirely 'accurate' may still embody 'bias' in the ethical sense of the word.

It is not sufficient for an algorithm to be technically 'right', it must also be ethically upright.



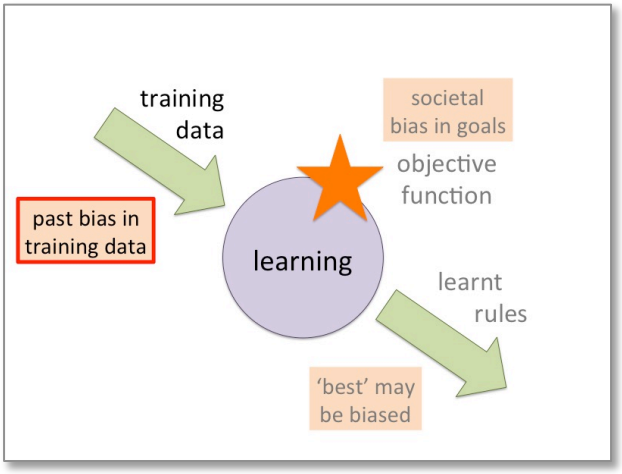
We'll structure our discussion of bias into three main sources based on the way they interact with the machine learning process.

The diagram shows a simplified view of machine learning. Training data is fed into the learning algorithm. The algorithm is guided by an objective or fitness function, which defines what it is to be a 'good' set of learnt rules. This all results in some form of learnt rules (where a 'rule' might consist thousands of weights in a deep neural network)

This gives rise to the three sources of bias:

1. bias in the training data from past biased human behaviour
2. bias in the goals from societal bias
3. even when bias in (1) and (2) are removed, the 'best' or accurate result may still be biased (in the ethical sense)

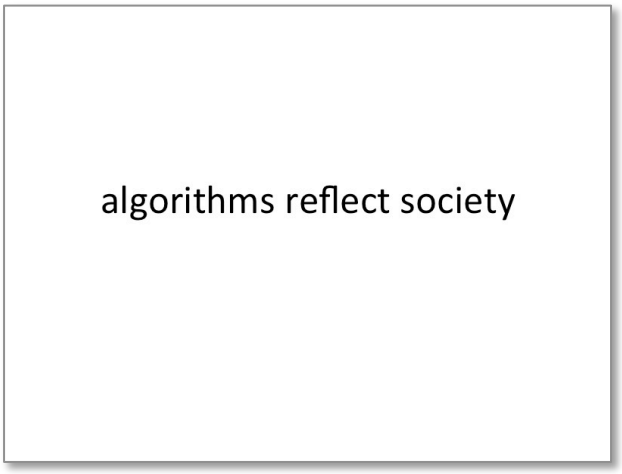
We will look at each in turn.



The first source of bias is in the training data.

If we want a machine learning system to distinguish cows from sheep then we feed it lots of images labelled 'cow' and 'sheep'; if we want it to recognise cancerous legions then we give it lots of labelled mammograms; if we want it to pre-select job applicants then we give it lots of CVs labelled as to whether or not they were called to interview.

However, if the person labelling pastoral images was confused by highland cows and labelled them 'sheep', then the ML system will confuse them too; if a certain form of cancer often got missed by radiologists, then the new cancer diagnosis system will also miss them; and if the past human selection of job applicants was racially biased, then the trained automated selection process will be similarly biased.



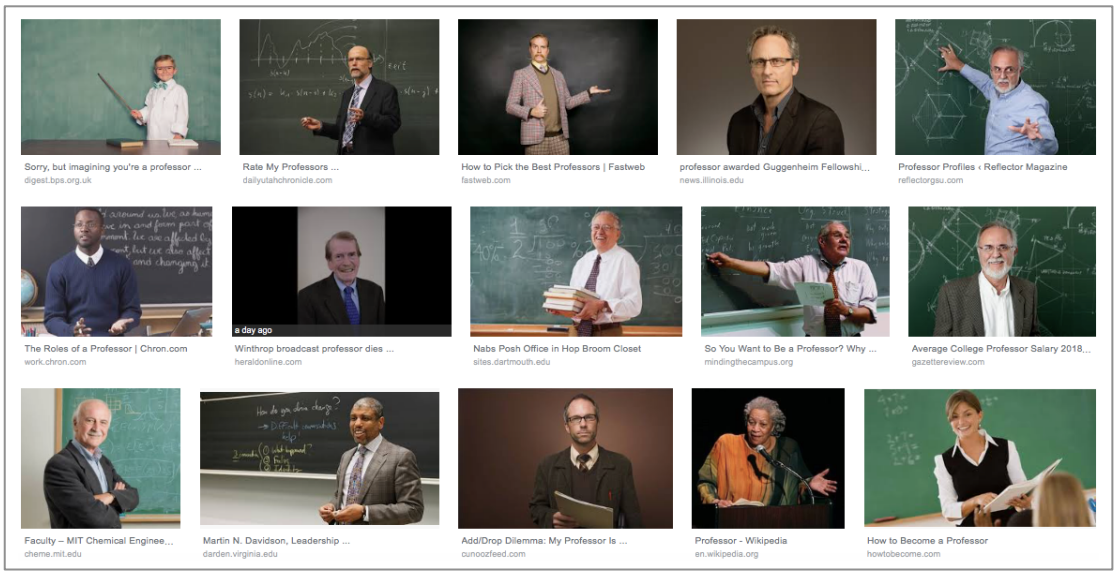
In general, the existing norms and biases of society will be embodied in past decisions and even special labelling for training. The machine learning system will faithfully copy the patterns of the training data and thus embody the self-same traits of society at large.

Happily this first source of bias is relatively well understood both in the technical literature and in the media. When Google processes billions of search queries; if people search for sexist or racist terms, this may naturally emerge as it autocompletes as you type. Similarly, an image search for "Professor" or "CEO" returns predominantly white male faces, but this precisely reflects the preponderance of such images in web pages labelled "professor" or "CEO".

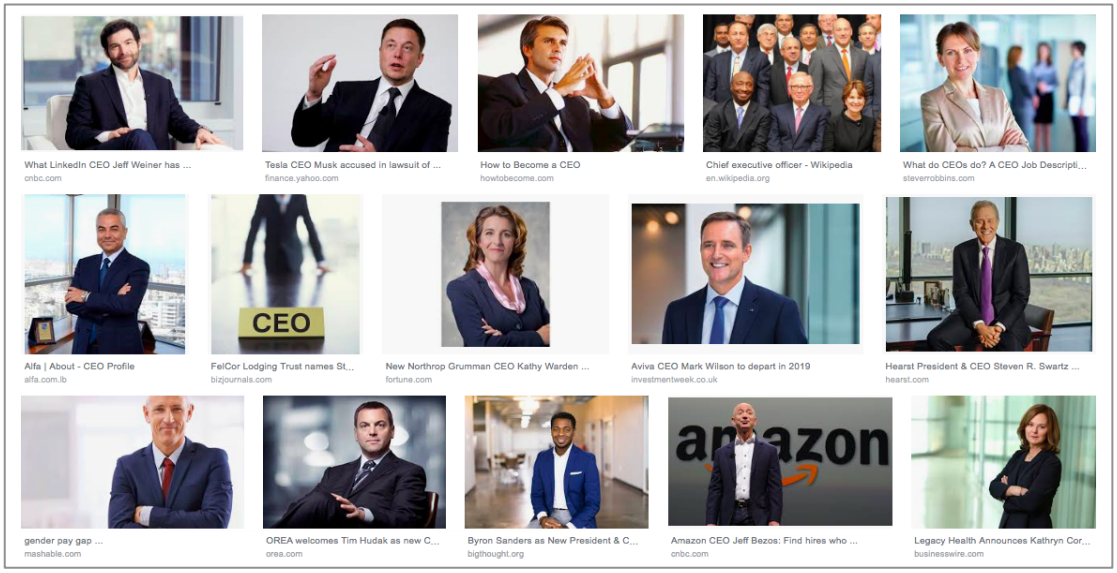
The last example is important as it is effectively reflecting the reality of society: senior positions are more likely to be held by white males.

Some courts in the US use automated systems to assess the risk of reoffending when considering sentencing or parole requests. These systems have been found to assess black offenders as significantly more likely to reoffend than white offenders even after balancing for other factors. However, if the police have been more assiduous in arresting and prosecuting black offenders, then, as a group, they will have a higher recorded reoffending rate, and therefore a statistically 'correct' system would reflect this. Of course the system would not have had the offenders colour as an explicit factor in the training data, but we shall see later that proxy' measures may effectively yield the same result.

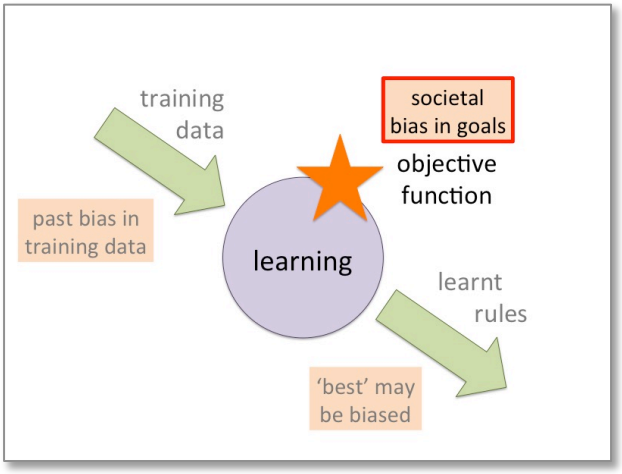
In these examples the training data reflects societal effects that are not neutral with regard to gender or race. It is no wonder the resulting systems exhibit bias.



Google image search “Professor” 13th Feb 2019 – 15 images, 10 white male



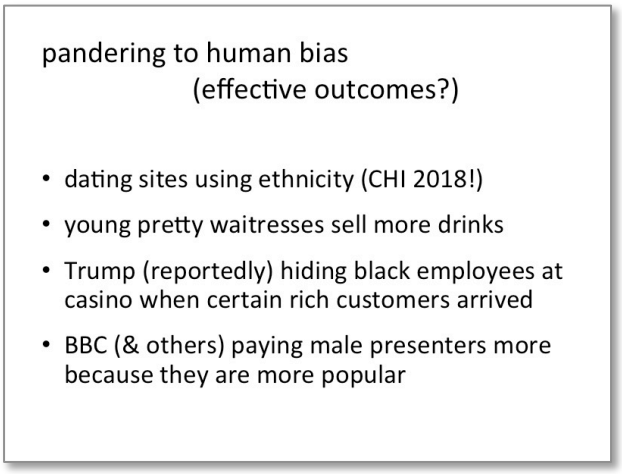
Google image search “CEO” 13th Feb 2019 – 14 individual images, 9 white male



The second source of bias is through what is called the objective or fitness function.

A development programme for autonomous cars might chose to train their system using a simulator. Initially the car would crash all the time, but gradually learn to be better. However to be 'better' the training system needs to know what 'better' is. If the measure of 'better' is minimising damage to the vehicle, then later, given a choice of a minor scrape against another car or mounting the pavement and ploughing down pedestrians, the car might chose the latter. If the measure of 'better; had been minimising fatalities and injury, then of course it would learn to make different choices.

This measure of 'better' is precisely the fitness function, and it is clear that this significantly affects the ultimate behaviour of the system.



In the previous example the objective function was explicit. However, often it is implicit, encoded in the preferences of society at large.

Examples of this are rife, not just in automated systems, but also human decision-making.

Consider a number of examples:

At CHI 2018, Chris Rudder, co-founder of the dating site OkCupid, gave one of the keynotes. Based on his book 'Dataclysm' [Ru14], he described how the data analytics exposed the choices of different genders, ages and ethnic groups. Much of this was unsurprising, but still shocking to see in raw numbers. However, more problematic was the way the dating site effectively pandered to these human biases. OkCupid was simply giving people what they wanted, maximising the chances they would find a profile they would like, but in doing so it made explicit choices, for example, to use ethnic profiles to determine who saw whom.

As a rule of thumb young pretty waitresses sell more drinks in a bar[1]. If you are a bar owner and have a number of candidates for a position, who would you choose to maximise sales? In the UK and many countries, it would be illegal to select based on two of the essential characteristics above (age and gender) although attractiveness is allowed. In UK universities and many companies, selecting based on physical appearance is also normally regarded as at least inappropriate if not against corporate rules.

In the 1990s, the Trump Plaza casino was fined $200,000 for deliberately moving black employees away from the tables when certain high stakes gamblers visited the casino [UP91]. Note that the casino had black employees, it was not being fined for discriminatory recruitment policies. The fine was because they were pandering to the racist whims of their customers.

In 2017, the BBC was widely criticised after it published pay gap figures showing that the most highly paid male presenters were receiving significantly more money than their female equivalents. When the story broke, the BBC Director General, Lord Hall was quoted as saying, "*The BBC does not exist in a market on its own where it can set the market rates. If we are to give the public what they want, then we have to pay for those great presenters and stars*." [BBC17]. Of course, if the public's perception of major presenters is biased, then that market 'value' will reflect this. In particular, experiments show that both male and female subjects harbour gender stereotypes that have changed relatively little in thirty years [HD16]. One such stereotype may lead viewers to unreasonably place more trust in male news presenters, thus creating the market demand and 'justifying' the BBC pay gap. However, the same argument could be made by Trump's casino managers or the shopkeeper choosing between candidates.

[1] I've tried to find definitive evidence for this common assumption, in case it too is simply a stereotype, but so far failed. However, there is an extensive literature of tipping behaviour and here there is solid evidence that tipping is influenced by the gender, race and 'attractiveness' of the waiter and the gender and race of customer [Mu12]. Crucially, on average, 'attractive females' are tipped more highly than other servers [LS00].

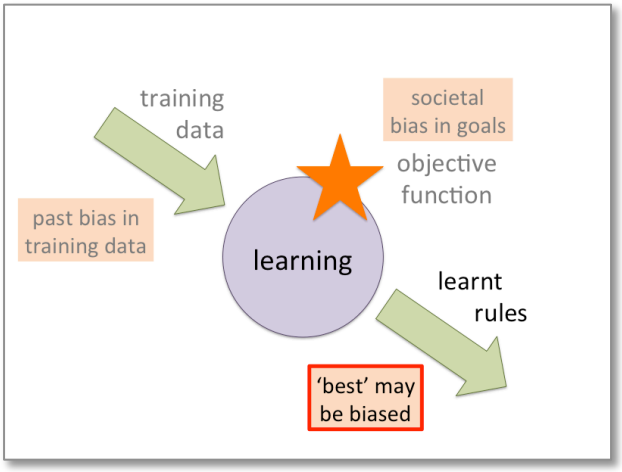
In each case the bias, prejudice and stereotypes of society mean that ‘good business’ would suggest making decisions that are driven by gender, ethnicity and other characteristics that would be deemed inappropriate, unethical or illegal if expressed explicitly.

However, imagine if in each case a machine learning algorithm or similar black-box technique was being driven by apparently neutral metrics such as popularity, consumer demand or tipping behaviour.

The only reason we know about OkCupid’s decision rules is that they obtained them in a two stage process, using data analytics to go from big data to comprehensible results, and then from that to hard-coded rules. If they had simply said, “*we put all our data into a recommender system*”, it would have been a very short, but uncontroversial, keynote. The cafe owner’s candidate selection program would ‘just happen’ to select attractive female applicants; and Trump Plaza’s shift allocation systems would ‘just happen’ to avoid black employees on the days certain customers were expected.

In some ways this is already happening with the BBC as the ‘market’ is the black box system.

Given an objective function to maximise profit, or audience share, the accurate and ‘best’ decision is ‘good’ business, but definitely not good.



Finally, even if the training data and objective are entirely unbiased, and the algorithms used have obtained the most accurate and optimal rules, the results of learning can still be 'biased' in the ethical sense.

For this part we'll focus on gender discrimination.

reinforcing societal/cultural norms

at school
- boys more likely to study STEM subjects
- girls more likely to study humanities

so, on average, with no other information

gender is an (albeit poor) predictor
- of communication skills
- and engineering knowledge

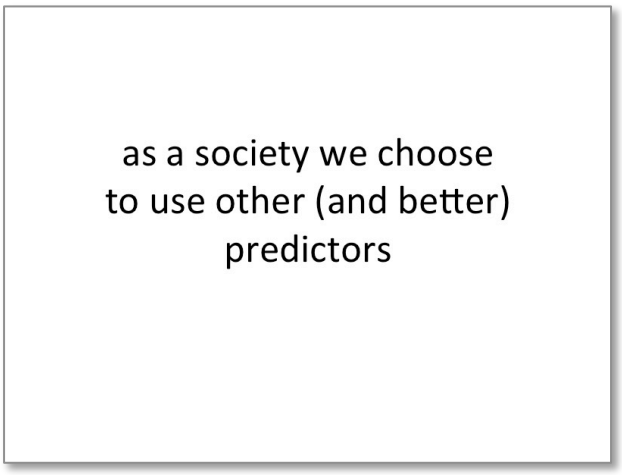
In most societies there are major differences on average between males and females, due to many factors, but most significantly the societal norms, expectations and sometimes explicit rules that influence our physical, intellectual and emotional development.

In the UK (and many countries), when there are choices in school, boys are more likely to take STEM subjects, such as chemistry, and girls humanities, such as history. This is clearly not an unviable fact of gender, as other countries do not have such a marked difference. However, in the UK, it is the case that, on average, girls and boys have had very different education by the time they leave school.

Now imagine you are selecting applicants for two jobs with the Antarctic Survey, one for a communication-rich role at Rothera research station on the Palmer Land peninsular, and the other for an engineering-related role at Davis research station near the Amery Ice Shelf, several thousand miles of ice and snow away.

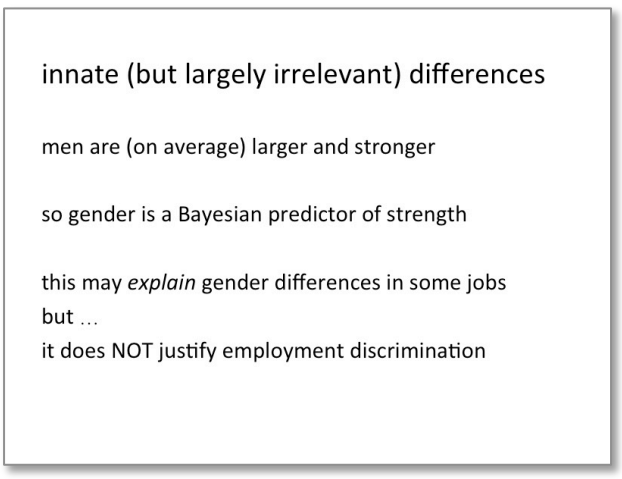
The applications are all in and you need to work quickly as the last ships, one for Rothera and one for Davis, are about to leave to get there before winter weather cuts off the bases for six months of dark winter.

You take the applications home and you have whittled the applicants down to two. Then, disaster, the dog eats the CVs. All you have left are the diversity information pages that you carefully separated and which contain information about gender, etc. There is no time to get fresh CVs and yet you must send the chosen applicants post haste to their respective ships. You peek at the forbidden diversity pages: one applicant is male and one female; one job is communications-rich, one engineering related … what do you do?



Because of our education system, gender is a predictor of communication and technical skills, albeit a poor one. The reason we do not use gender as a predictor is not because it lacks predictive power; instead it is because we *choose* not to. It is an *ethical* decision.

As a society we choose to use other (and actually far better) predictors. We may look at exam result, or run our own tests that more directly assess the skills or knowledge we require, but we choose not to use gender irrespective of whether it offers any predictive power.



This even extends to innate differences. There is much discussion about this, for example, whether gender differences in spatial cognition are due to innate neurological differences, or simply the effects of early social conditioning. For most purposes whether these differences are genetic or environmental is immaterial. Crucially most are about small differences in average behaviour compared with much larger inter-personal variation.

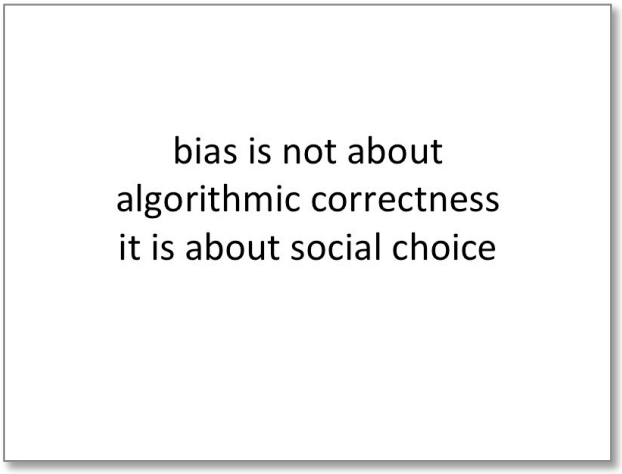
One of the most substantial differences is that men are, on average, larger and stronger than women. That is gender is a predictor of strength. Of course individuals of both sexes vary in strength, but this can be used as a rule of thumb.

This difference in average strength may *explain* observed differences in the workplace, for example, trades that involve frequent heavy lifting may end up with more men working in them. Of course it would *not* justify employment discrimination.

However, now imagine a very socially conscious building company. They are very traditional in terms of methods (a lot of heavy lifting), but advanced in the use of IT, so they decide to create an automated system to help with hiring. In order to avoid bias in the system, they conduct an extensive experiment. One thousand people are recruited 50% male, 50% female and employed for 2 months, with their productivity heavily monitored. At the end of the experiment a machine learning system is given the measured productivity of subjects together with their CVs and builds a predicator of productivity. Being an ethical company, the gender and other protected characteristics would be removed from the CVs before they are entered into the learning system.

If the system were entirely opaque one would just have to trust it. The entire process was gender-blind, so surely the resulting system would be unbiased?

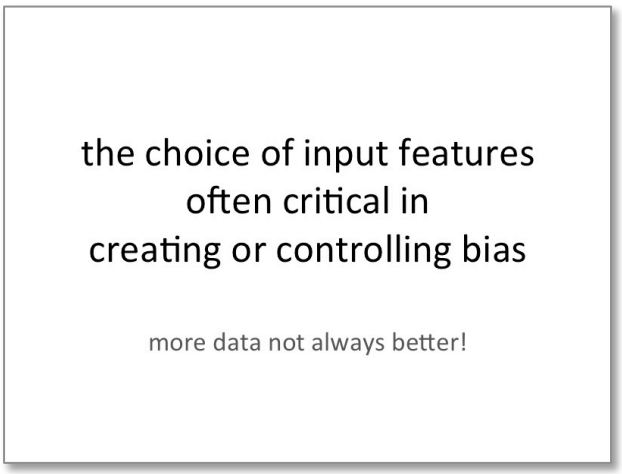
Now imagine a learning system that creates more transparent rules. You start to interrogate it and find that school exam subjects are being used and the rules effectively say, “if the person has taken STEM subjects then hire them”. Now STEM subjects at school are almost certainly not useful on an old-fashioned building site, but they are a proxy indicator of gender, which in turn is a crude predictor of strength.



Note here that the algorithm is producing the 'best', most accurate estimator given the data available. However, bias, in an ethical and legal sense, is not about algorithmic correctness, it is about social choice.

Note also that if the job had required technical ability or good communications, then exam grades would be deemed a reasonable and acceptable decision criteria. The exam results would correlate with gender, but would be directly relevant to the job. The problem in the building site is that they are not clearly relevant to the job at hand and merely acting as *proxy* gender measure.

In other words, exactly the same training data could yield ethical or unethical (and legal or illegal) outcomes depending on context.

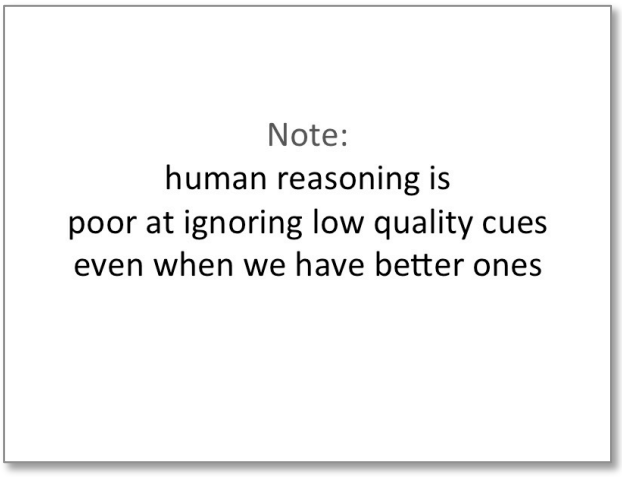


Note too that in this and other examples, the choice of input features being fed into a learning system is often critical in creating or controlling bias. The most obvious is deliberately not including explicit gender or similar indicators, but as we have seen this is *not sufficient*.

In the case of the builder, the problem is partly that if there is no direct measure of physical strength on the CVs, then the system will choose the 'next best' thing and may latch onto (proxy measures of) gender.

A strong guard against this is to make sure you collect relevant features. If the system has a good measure of physical strength, it is less likely to fall back on gender, which is a relatively poor predictor.

Furthermore, as well as removing explicit gender indicators, one might consider also deliberately not including what appear to be irrelevant features. This may have technical benefits reducing over-learning, but also make the system less likely to have potential proxies to latch onto.

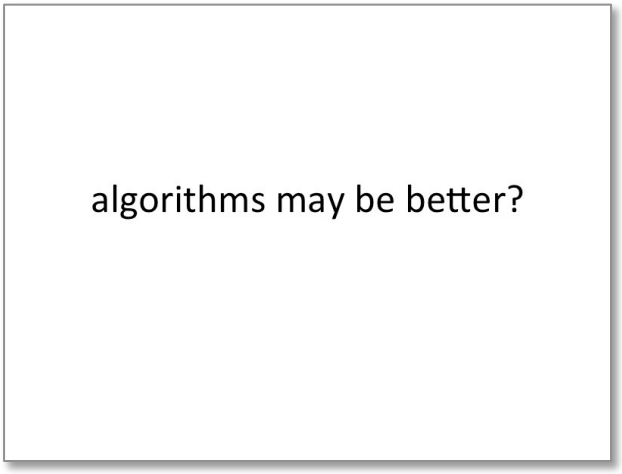


The examples point out potential dangers for machine learning systems. However, anti-discrimination legislation predated the widespread use of AI. The legislation exists precisely because humans haven't done so well at these issues prior to automation.

The human perceptual and cognitive system has developed primarily for information poor environments, where you have to make the most effective inferences from scant data. Now we live in a world of information overload, but with the same perceptual and cognitive system as our cave-dwelling ancestors.

Crucially we are poor at ignoring low quality cues even when there are better ones to hand.

In one example of this, Salmoni investigated people's ability to assess the quality of search results based on the title and snippet as commonly found in web search results [Sa04], Showing the title only yielded better relevance than the snippet only; however, when they were sown together, rather than being better still (more information), the effectiveness fell between the two on their own. Even though the snippet was not contributing to the subjects' ability to assess relevance, they were unable to ignore it and hence performed less well than if they had had the title alone.

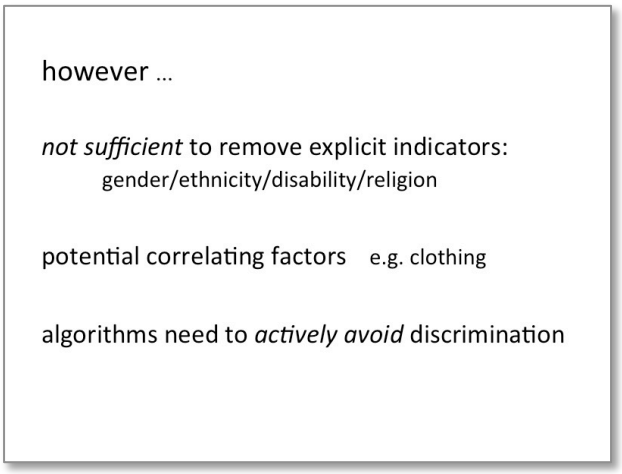


It may even be that algorithms could be better.

A study was performed some years ago of people admitted to hospital for heart attacks. The doctors gathered many test results and other forms of evidence and used this to decide amongst a few different forms of treatment. Retrospective data was then collected including the original diagnostic features and the clinical outcomes for the patients after a few months, whether they had recurrence, or indeed died.

The data was used to train a classifier, however this was not used as an automatic diagnosis system to replace the doctors' judgement. Instead the analysts examined the rules created by the system and realised that the optimal classification depended on four features only.

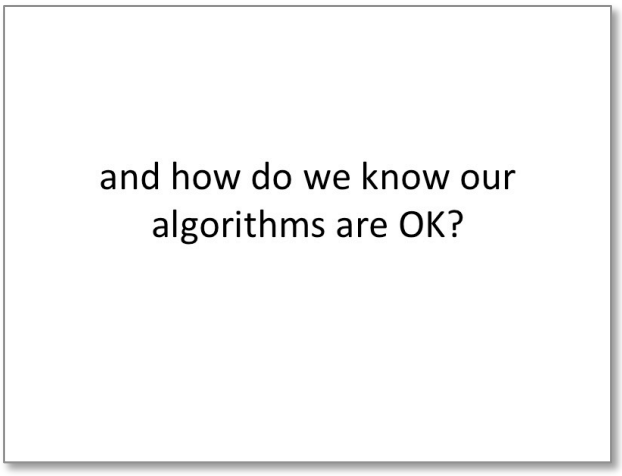
The doctors were told about this and changed their clinical practice: only collecting the four relevant factors, but otherwise using their clinical judgement as before. They found that their own clinical outcomes improved. By not collecting data and never seeing it, they became better at their job



While in principle algorithms could behave better than people, the reality is still far from this.

Crucially, as we have seen, it is not sufficient to remove explicit indicators of gender, ethnicity, disability, religion, or other protected characteristics. Many discrimination cases relate to indirect forms of discrimination, for example, demanding a particular headwear when this is not essential to the job, which effectively discriminates against Sikhs or those wearing the hijab. As we have seen it is easy for a machine learning system to accidently latch onto proxy measures of a protected characteristic.

Instead, algorithms need to *actively avoid* discrimination. For example, after training an algorithm on gender-blinded data, one could deliberately re-introduce gender and build a causal model. If the impact of features on the final decision is factored through gender, then that is a good indication that the features were acting as a proxy for gender. One could even imagine building this into the original learning process.



Whether or not algorithms are better or worse then humans at making ethically unbiased decisions, how do we know?

We might look to external audits of the statistics comparing the way different groups are dealt with by a system or process, whether by human or machine. A good example of this is the way many companies now publish pay gap data.

Note that such external statistics do not answer the question "is my process biased", but do offer evidence to pursue and investigate in more detail.

For example, some years ago gender pay-gap data was published for software development employees of a large number of US hi-tech companies. In many cases the company broke down the data by job role, giving quite detailed data across the sector. As expected, the proportion of female employees was small, but this could be explained by lower numbers of women taking computing-oriented degree subjects. More significantly, in all but a small number of companies there was a pay gap with male employees paid more on average then female counterparts in equivalent roles. This was the main headline leading to the data being widely publicised and shared on social media.

However, the data also included average time in post, also broken down by gender. Typically the average time in post was also shorter for women. In all but four cases, this was a perfect predictor of the pay gap: if men had longer than average time in post for a specific company and role, then they had higher average salary, if the women had longer time in post, they had higher average salary. The real question to ask was not about the pay gap, but about *retention*. Why the difference of time in post? As this persisted at different levels of seniority it was not simply differences in past employment practices, nor due to maternity gaps. It could be that women were promoted faster, or it may be that women left because of the 'boys club' atmosphere of hi-tech companies.

Of course, it could also have been the other way round, no apparent pay gap, but, say, hidden bias in underlying recruitment. Equal pay for all in the same post, but only after having a higher recruitment bar for some group. This is perhaps what one might expect to see in lower paid jobs such as in hospitality.

The stats are the beginning, not the end of an investigation into bias. It is not sufficient to look at the overall numbers, but we must dig into the reasons that led to them.

Bias is not the only reason we need to dig more deeply into algorithmic (or other) decisions.

- ***safety*** – When an autonomous car has an accident, we need to understand what went wrong in order to prevent similar future accidents. The airline industry has long-standing rigorous methods for this adopting a forensic analysis of every accident. Normally car accidents are not treated with the same level of detail even though in total they cause a far greater death toll. This is in part because they are each individually smaller, but also because it is too easy to blame the driver: human error. However, software-controlled cars will mean that causes of accidents will be more likely to have repeatable causes.
- ***democracy*** – The past few years have seen great worries about the ways algorithms potentially undermine democracy. Sometimes this is about

deliberate practices such as the Cambridge Analytica scandal, or twitter-bots. However, perhaps more worrying is the way that the "what people want' (objective function) algorithms of search engines and social media create bubbles of like-minded information, allowing us to each feel we are in the fact-based majority against an ignorant, albeit vocal, minority.

- ***health and wellbeing*** – Imagine you are the senior executive of a soft drinks manufacturer that wishes to adopt an ethical advertising policy. You do not deliberately advertise in children's magazines or on children's TV, but how do you know whether your online advertising, which may be driven by keywords or much more complex algorithmic mechanisms are not implicitly targeting children?

- ***social issues*** – Not long before the talk in 1991, I was told by my bank that I would need to pick up chequebooks directly from my branch; they couldn't post them to me because I lived in a 'high-risk postal code'. In other words they did not trust the honesty of my neighbours if it were misdelivered to the wrong house … and, by implication, they would not trust me if they were considering posting a chequebook to a neighbour! This was a minor inconvenience, but the cost of everything from car insurance to interest rates on loans themselves are driven by a wide variety of factors including the area you live, often linked directly or indirectly to your socio-economic status. At one level this is simply reflecting the market, but of course the same could be said about some of other discriminatory effects we have discussed. Unless we understand how these decisions are made it is hard to assess their ethical status.

- ***science*** – There are similar worries in the scientific community that big data approaches to science may well be 'discovering' relationships that later turn out to be spurious [Gh18]. Bias can also creep into the most apparently 'objective' basic science. Henrich et al. [HH10] have pointed out that most cognitive psychology has been developed using experimental subjects that are WEIRD (Western, Educated, Industrialized, Rich, and Democratic). Their meta-study showed how what appear to be fundamental cognitive and perceptual phenomena, such as the Müller-Lyer illusion, are often culturally determined

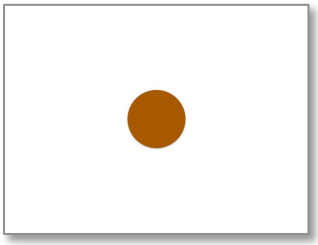
In general, for many kinds of algorithms and complex rule-driven human processes, we need to be able to ask the question “why?”

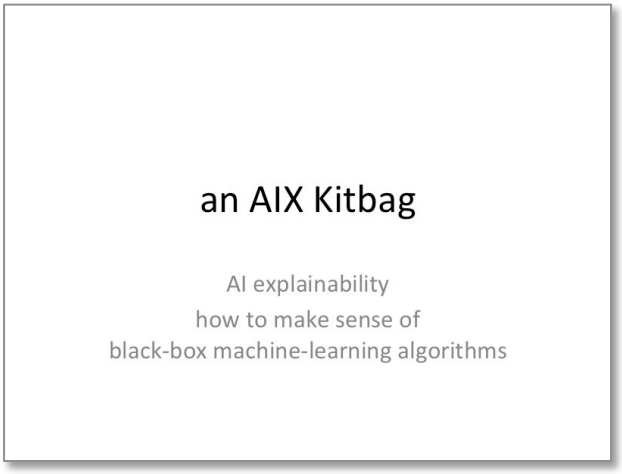
Why did that car crash?

Why was I refused a loan?

Why did the police stop me in the street to question me rather than all the others walking by?

This emphasises the need for some form of transparency or explainability in complex algorithms.



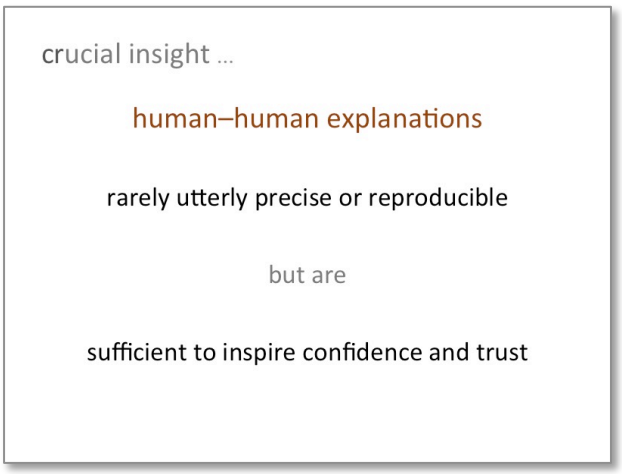


The field of explainable AI has been growing rapidly over recent years in the face of these issues.

Some suggest that deeply opaque methods such as deep learning are by their nature unexplainable. However, there has also been promising work, both in more traditional symbolic AI (e.g. argumentation-based reasoning) and in sub-symbolic AI and machine learning (e.g. hot-spot analysis of critical regions for image-recognition systems).

Often results are very specific to a domain or technique, but it is evident that some of these offer potential methods that could be adapted or core principles extracted so that they could be used more widely.

With colleagues I have been developing an AI explainability Kitbag (AIX Kitbag), which collates high-level heuristics that can be reapplied to different underlying algorithms and domains [Dx17].



Crucial to this endeavour is the insight that human–human explanations are rarely utterly precise or reproducible.

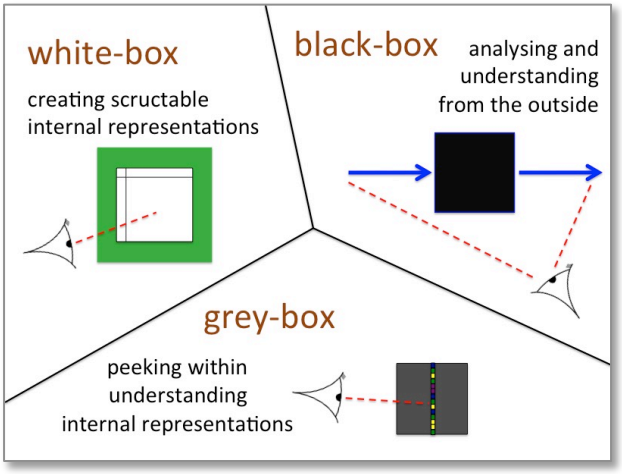
If at a restaurant you were asked why you chose a particular main course you might say something like, "well I usually go for a steak, but it was late and I wanted something lighter; I'd had fish last night, so chose a salad." Within this are many vague concepts and open questions. Why normally choose steak? What do you mean by 'lighter'? Why not have fish two nights running? However, for most purposes this would be a sufficient explanation. Of course the statement might elicit further questions, "why didn't you go for the spinach brûlée, I know it sounds odd but is actually quite delicious?" Of course, both questions and answers themselves might leave aspects only roughly defined, but sufficient for a discussion about food.

We do not try to explain in terms of the firing of individual neurons in our brain, or try to make precise every nuance. Furthermore, the explanations we provide are often rational reconstructions, ways of make sense to ourselves, as much as to others, the complex interweaving of conscious and unconscious processes in our minds.

In human–human discourse statements and explanations are part of a process of mutual understanding that enables further action or communication. Studies repeatedly show an incremental processes of unfolding of partial statements rather than precise detailed monologues (except in the university lecture theatre). For example Grices's *conversational maxims* include to "make your contribution as informative as is required" – but no more [Gr75] and Clark and Brennan suggest that our conversational utterances will perforce involve levels of ambiguity, which are confirmed or disconfirmed as part of on-going discourse, with the aim of creating a sufficient *common ground* of understanding for future conversation and action [CB91].

In short, the purpose of an explanation is to inspire confidence and trust to allow future mutual action, or possibly to create sufficient openness to allow critique or dispute.

When we look at machine–human explanations in this light it is often easier to see how we may at least make complex big-data analysis, deep learning and similar algorithms comprehensible if not utterly 'explained' to the last possible detail. Indeed such an over-detailed explanation, would probably, for the human, be no explanation at all.

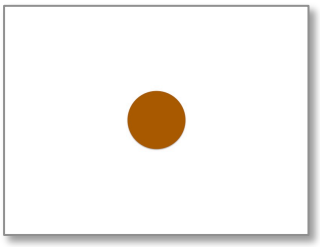


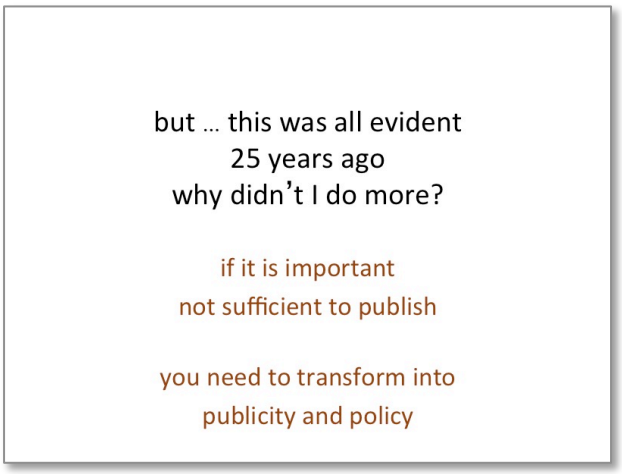
The current AIX Kitbag divides explanations into three main categories, each of which, at present, has five or six heuristics.

white-box techniques – These are algorithms which by their nature or with minor modifications naturally have understandable internal representations. The choice of decision trees in Query-by-Browsing is because these were relatively easy to understand and could easily be transformed into standard database queries. Note that, as in the genetic algorithm version of QbB, it may be sufficient for the decision rules to be explainable, whether or not the process that created them is. Another example is that in the early days of neural networks, there was work on 'hardening' the sigma activation functions used during back-propagation, and turning them into simple thresholds, often resulting in more comprehensible (albeit large) Boolean networks.

black-box techniques – Here one treats the process is a black-box, but attempts to make sense of it from the outside. If the police have suspected terrorists under surveillance, they will not walk up to them and ask, "why are you buying fertiliser?". Instead, they will attempt to determine plans, motives and reasons based on the *observable behaviour* of the suspects. With an algorithm or machine learning systems, this is possibly easier as we can run 'experiments' on the algorithm. For example, one can perform sensitivity analysis, tweaking the inputs slightly to see whether they change the output of the algorithm. Where small changes in inputs create large changes in the output (e.g. a different classification), then they are clearly critical in the internal process.

grey-box techniques – Where the internal process has some sort of intermediate representation, such as one of the internal layers in a multi-layer deep-learning network, this can be used to look for black-box explanations in both directions. Typically the lower levels, closer to the input, will be framing broad conceptual categories, or performing a form of dimensional reduction. Here one might use techniques to help a human say something like "that cluster of nodes is about whether there is cat in the scene", or "that is about food being spicy". The upper levels may be amenable to transformation onto a more logical/symbolic representation, so that together the explanation might be "choose the meal if it is spicy, but not too expensive".





Much of this talk was evident in the earlier talk and book chapter 25 years ago. Indeed, even some parts of the AIX Kitbag could have been written then.

So, given this was evidently a problem at that point and there were potential ways to address it, did I not do more then?

To some extent it may have been ahead of its time, but the truth is I never tried to do more. I was aware that these were important issues, and I also knew potentially promising directions, but apart from a small amount of follow-up work in Query-by-Browsing, I did nothing.

Why?

The answer is that I had done my academic job: published the work, put it in the public domain, told the world, ticked the box and so I moved on to the next problem.

At the time I thought academics who tried to push their work, were self-publicists, or simply lacked many ideas. It was only years later, I realised the, quite obvious, truth, that if something is important, whether academically or societally, then it is important to not just tell people once in a single publication, but make sure they heard.

It is not sufficient to publish, but also the work needs to be publicised, and in the case of socially important issues, to push this into practical systems that work for people or into public policy.

References

[Ba15] Barrett, B. (2015). Google Maps Is Racist Because The Internet Is Racist, Wired, May 23, 2015. https://www.wired.com/2015/05/google-maps-racist/

[BBC17] BBC (2017). BBC pay: Male stars earn more than female talent. BBC News, 19 July 2017. https://www.bbc.co.uk/news/entertainment-arts-40661179

[Bo19] Bob, Y. (2019). Ex-Gov't Agent: Crisis Worse Than 9/11 Could Emerge From AI Arms Race. Jerusalem Post, Feburary 12, 2019. https://www.jpost.com/Israel-News/Ex-govt-agent-Crisis-worse-than-911-could-come-out-of-AI-arms-race-580459

[Ca15] Carpenter, J. (2015). Google's Algorithm Shows Prestigious Job Ads To Men, But Not To Women. Independent, 7th July 2015 https://www.independent.co.uk/life-style/gadgets-and-tech/news/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-10372166.html

[CAB13] Citizens Advice Bureau (2013) 22% don't have basic banking services needed to deal with Universal Credit. Retrieved 29/01/2014, from http://www.citizensadvice.org.uk/index/pressoffice/press_index/press_office20131105.htm

[CB91] Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine, & S. D. Behrend (Eds.), Perspectives on socially shared cognition (pp. 127–149). Washington, DC: American Psychological Association.

[CEU16] Council of the European Union (2016). Position of the council on general data protection regulation. 8 April 2016. http://www.europarl.europa.eu/sed/doc/news/document/CONS_CONS(2016)05418(REV1)_EN.docx.

[Cr19] Andy Crabtree (2019). Right to an Explanation Considered Harmful. Computational Foundry Seminar, Swansea University, 19th February 2019.

[Cy91] Cypher, A. (1991). EAGER: programming repetitive tasks by example. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '91), Scott P. Robertson, Gary M. Olson, and Judith S. Olson (Eds.). ACM, New York, NY, USA, 33-39. DOI:10.1145/108844.108850

[DT15] Datta, A., Tschantz, M., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings, Proceedings on Privacy Enhancing Technologies, 2015(1), 92-112. doi: https://doi.org/10.1515/popets-2015-0007

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

[DP94] A. Dix and A. Patrick (1994). Query By Browsing. Proceedings of IDS'94: The 2nd International Workshop on User Interfaces to Databases, Ed. P. Sawyer. Lancaster, UK, Springer Verlag. 236-248.

[Dx17] Dix, A. (2017). An AIX Kitbag: Discussion Document), version 1. (unpublished). Released 21/12/2017 (Minor revisions 12/5/2018). https://alandix.com/academic/talks/sufficient-reason-2018/AIX-kitbag-v1b.pdf

[Dx18] Dix, A. (2018). Sufficient Reason. Keynote at HCD for Intelligent Environments, BHCI, Belfast, 3rd July 2018. http://alandix.com/academic/talks/sufficient-reason-2018/

[Dx98] Dix, A. (1998). Interactive Querying - locating and discovering information Second Workshop on Information Retrieval and Human Computer Interaction, Glasgow, 11th September 1998.

[Gh18] Gholipour, B.(2018). We Need to Open the AI Black Box Before It's Too Late: If we don't, the biases of our past could dictate our future. Futurism, January 18th 2018. https://futurism.com/ai-bias-black-box/

[Gh18] Ghosh, P. (2018). AAAS: Machine learning 'causing science crisis'. BBC News, 16 February 2019. https://www.bbc.co.uk/news/science-environment-47267081

[Gr91] Greake, E. (1991). Neural network keeps fusion plasma in shape. New Scientist, page 27, 12th October 1991.

[Gr75] Grice, H. P. (1975) 'Logic and conversation'. In P. Cole and J. Morgan (eds) Studies in Syntax and Semantics III: Speech Acts, New York: Academic Press, pp. 183-98.

[Gu16] Guardian (2016) Microsoft 'deeply sorry' for racist and sexist tweets by AI chatbot. The Guardian, 26th March 2016. https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot

[HD16] Haines, E. L., Deaux, K., & Lofaro, N. (2016). The Times They Are a-Changing … or Are They Not? A Comparison of Gender Stereotypes, 1983–2014. Psychology of Women Quarterly, 40(3), 353–363. https://doi.org/10.1177/0361684316634081

[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

[HM15] Hawking S, Musk E, Wozniak S et al (2015) Autonomous Weapons: an Open Letter from AI & Robotics Researchers. Future of Life Institute. http://futureoflife.org/AI/open_letter_autonomous_weapons

[HH10] Henrich J, Heine S, Norenzayan A (2010) The weirdest people in the world? Behavioral and Brain Sciences, 33(2-3):61–83. doi: 10.1017/S0140525X0999152X

[Ku13] Kushner, D. (2013). The Real Story of Stuxnet. IEEE Spectrum, vol. 50, no. 3, pp. 48-53, March 2013. DOI: 10.1109/MSPEC.2013.6471059

[La18] Lapowsky, I. (2018). Google Autocomplete Still Makes Vile Suggestions. Wired, 2nd Dec 2018. https://www.wired.com/story/google-autocomplete-vile-suggestions/

[Le82] Leith, P. (1982). "ELl: An expert legislative consultant", in Proceedings lEE Conference on Man/Machine Systems, Manchester, 1982.

[Le85] Leith, P. (1985). Legal knowledge engineering: Computing, logic and law. PhD thesis Open University. http://oro.open.ac.uk/56914/1/371041.pdf

[LW18] Levin, S. and Wong, J. (2018). Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian. The Guardian, 19th March 2018. https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe

[LS00] Lynn, M & Simons, T (2000). “Predictors of Male and Female Servers’ Average Tip Earnings.” Journal of Applied Social Psychology 30: pp. 241-252.

[MD14] A. Morgan, A. Dix, M. Phillips and C. House (2014). Blue sky thinking meets green field usability: can mobile internet software engineering bridge the rural divide? Local Economy, September–November 2014. 29(6–7):750–761. (Published online August 21, 2014). doi: 10.1177/0269094214548399

[Mu12] Mulinari, P. (2012). What is that little extra? Exploring the social norms of tipping. International Labour Process Conference, ILPC 2012. https://www.ilpc.org.uk/Portals/56/ilpc2012-paperupload/ILPC2012paper-finalpaperilpc2012_20120326_101415.pdf

[No18] Noble , S. (2018). Google Has a Striking History of Bias Against Black Girls. Tim, March 26, 2018. http://time.com/5209144/google-search-engine-algorithm-bias-racism/

[Qu86] Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1):81-106

[Ru14] Rudder, C. (2014). Dataclysm: Who We Are (When We Think No One's Looking). Fourth Estate. ISBN: 0008101000

[Sa04] Salmoni , A. (2004). Task-based Judgements of Search Engine Summaries, and Negative Information Scent. PhD Thisis, Cardff University, September 2004

[UP91] UPI (1991). Trump Plaza fined $200,000 for discrimination. UPI Archives, June 6, 1991. https://www.upi.com/Archives/1991/06/06/Trump-Plaza-fined-200000-for-discrimination/2869676180800/

[Vi13] Vincent, J. (2013). Russian Nuclear Power Plant Infected By Stuxnet Malware Says Cyber-Security Expert. The Independent, Tuesday 12 November 2013. https://www.independent.co.uk/life-style/gadgets-and-tech/news/russian-nuclear-power-plant-infected-by-stuxnet-malware-says-cyber-security-expert-8935529.html