Data Ethics, AI and Responsible Innovation

Nick Hood

16 November 2020

Contents

Introduction

These are my notes from participation in the Edinburgh Data Ethics MOOC, running in November and December 2020. Any errors here are my own responsibility.

Nick Hood
@NixImagery

This document was last updated on 16 November 2020 at 11:32.

Course overview

Week 1 Ethics and Law What are ethical values? Can we rely on legal regulation? What are the most pressing issues facing data driven industries?

Week 2 Crime and Justice Should we use predictive policing and sentencing algorithms? How can biases sneak into such algorithms? How can we remove them?

Week 3 Home and City What are the promises of smart homes and cities? How can they impact our privacy and freedom? How can we design them to protect those values?

Week 4 Money and Markets Can future tech lead to a world without money? Can algorithms help us fairly distribute resources? How could we design a fair AI?

Week 5 Life and Health Should we keep genetic databanks? Would you trust an AI doctor? What are the principles of responsible research and innovation?

Week 1: Ethics and Law

This starts with some housekeeping, introductions to the tutors and setting up a profile within the course. Dr. Ewa Luger gets us started by considering a "broad overview" of **ethics**, using a radio interview about the Cambridge Analytica (CA) scandal as a resource to get us thinking before we embark on a more detailed introduction to ethics.

- Ethical 'Issues'
- Introduction to Ethics
- Legal and ethical Considerations
- Information, Control and Power

Ethical 'Issues'

In the clip, problematised issues with the CA matter included:

- the way data was gathered from people
- the fact that the data was then passed on to another party
- that the data was now available for commercial use
- that the data was used for political purpose
- the FaceBook (FB) login on an app gave permissions for all of users' FB data to be used and shared
- people aren't outraged, or even bothered by this enough to take action

Discussion forum task Think back over your professional life and identify one example of an ethical issue that you personally experienced. In the discussion forum, write a brief summary (around 50-100 words) describing this issue - try to explain the context, what happened, and why you felt it was an ethical issue. After you've done this, comment on two other posts.

I had to take a quick look at the forum first to gauge the level of ethical matters that were being shared. There seems to be a discernible difference between **moral**, **legal** and **ethical** matters in the revelations on the forum. My own example came from quite a selection of professional experience that ranged from legal but immoral, moral but illegal, unethical but moral and legal. I picked an example from a while back.

> I was project manager in a business that made flight crew training simulators. These devices were multi-million-dollar complex machines that used specially designed and manufactured circuit boards mounted in standard racks to drive the various components of the device. I had an internal review one morning that failed badly when the equipment that had been working well the night before suddenly failed to function. It was extremely embarrassing for my team. We discovered later that day that in the night, another project manager had switched a key circuit board for a faulty one from his own machine.

The other examples I commented on or read seemed to me to be not so much ethical issues as difficult choices or situations. This is causing me to wonder if I have a secure understanding of the term, "ethical". Hopefully the next section will clear that up for me.

The challenge of data-driven innovation

Issues arising from the 2018 Cambridge Analaytica scandal are identified as:

- the value derived from data is not evenly distributed
- power rests in the hands of very few
- the potential for harm and inequality is high
- the potential impact of bad actors

These are, in my view, issues of power arising from wealth inequity and not particularly related to anything technological. The difficulty is that technology amplifies the effects (as it amplifies inequities in a classroom – this much we have learned in moving to "Digital First" pedagogies resulting from the COVID measures).

The basic principles of ethics are described as, "how we should address these issues, whilst minimising harm, ensuring morally good outcomes and fairness, and protecting human autonomy."

Data at the heart of it

For the purposes of the course, this definition of ethics is set in the context of *data-driven innovation* which makes use of large amounts of data to train algorithms that make predictions or provide insights for decision making. Machines are increasingly now involved in the decision making too, and in taking actions. These machines are described as *intelligent*.

Visibility of rationale

These machines include, of course, neural networks and we know from the way that these pattern-matchers work, that other than the input and output layers, it is not possible to determine the path of the rationale – the specific choices ot selections being made in the intermediate steps – that underlies the outputs.

Ambient intelligent systems

Interactional moments are perceived to provide friction in the operation of systems, therefore designers are removing these for a smoother experience – think of entering your password every time you wanted to look at your phone, now replaced with fingerprint or face recognition. The result is that such interactions become invisible to the user: it is the environment we interact with in a natural way. Unnaturally, *the environment is reading us to discern intent.*

Human –> AI interaction

This is problematised in a video clip, suggesting that this interaction is not yet defined. I rather think that Turing (1950) had a clear enough view of what it might look like, but recent innovations have increased *system opacity* such that understanding how the system makes choices based upon how it perceives us is difficult. In some sense, we are already familiar with this idea with our less tech-savvy relatives who need support working with machines. It gets harder when you don't know there's a machine.

Utility

Comparison with basic utilities like water, power, etc., is made from a very privileged first-world perspective: *we only notice them when they fail.* These systems were unregulated in their early days, too, and some of the horrors of human behaviour (Edison, for example) are long forgotten. As data-driven innovation moves to infect our lives by becoming a utility, the question is asked, "should it be regulated?" First steps have been made in this, of course, and we see the return of interactional friction with things like the cookie quiz on GDPR-compliant websites. The monetisation of data has led to the term, "surveillance capitalism", which describes how our identity and behaviours have become the product in a high-value industry.

Policy lag

Clearly, policy makers are well behind the curve when it comes to data-driven innovation. GDPR and similar legislation have added friction yet advances are made at speed. The additional reading (Wagner, 2019) is critical of policy makers and suggests that ethics is what corporations do to avoid government interference.

Values

What we value in society is a cultural attribute. Values are therefore dependent on context and are not absolute: I have often said that every principle has a price attached to it, and that if the price is right, all principles are for sale. This has caused strong reaction when vocalised that way, especially from those who feel that they themselves hold high principles in comparison with others. Clearly, passions are aroused when deeply-held values are challenged, as the Guardian article on Peter Seeger illustrates.

Examples are discussed in the course material to illustrate the point that values depend very much on context. This statement is a call to action, perhaps, talking of moral dilemmas and conflicting values:

> "Our job, as ethical actors, is to identify what they are, and then negotiate how best to instantiate and balance them."

Research is as robust as you'd expect for the social sciences, with this, from a study of a small sample of people ($N < 700$) in three countries with similar (colonial) cultural pedigree:

> "... people endorse the same values to a similar extent across countries and also instantiate them similarly." (Bardi et al., 2018)

Not much of a surprise there, given the evolutionary path of the handful of people studied.

Case studies

Proejct Maven and the Google employee revolt against it is presented as the first case study of how corporations can make choices that do not align with the values of the stakeholders, in this case, the staff. That project sat in stark contrast to Google's "Don't be evil" mantra, still enshrined in its code of conduct (cf. the characterisation of ethics as an escape from regulation by Wagner (2019)). The military application of data-driven innovation is clearly compelling, and others are emerging, including Clearview AI's face recognition app, which the company has worked hard to present as ethical use of technology.

What are ethics?

A branch of moral philosophy: a codification of habits that are valued within the context of a culture: ergo, no right or wrong answers, just human custom. *Normative* ethics are those that define what we *ought* to do, without concerning themselves with what we *actually* do. Within that class of ethics are 3 persepctives on how we should act:

Deontological ethics (rules) rules of duty and obligation – perhaps universally agreed upon, like not killing each other, or hurting animals. Some things are always right, others, always wrong, no matter the consequences.

Teleological ethics (consequences) focus on the outcome of actions – actions for the greater good, perhaps, or the end justifying the means. *Consequentialism* is the view that the moral quality of a choice is decided solely by its outcome.

Virtue ethics are about judgement of people's character or moral fibre.

The obvious issue here is that life is complicated, isn't it? We can't make a simple set of rules because the complexity of life requires us to make choices sometimes that go against those simple rules for a better end. Is there any point, then, in having rules in the first place? Well, clearly, because we want everyone else to make choices that don't harm or disadvantage us.

The trolley problem

Scenario 1 A runaway train is travelling on a railway track towards 5 people. You can't warn them, but there is a lever that you can operate to switch the tracks. The problem is that there is a person on the other track. **Do you pull the lever or not?**

Scenario 2 As above, but now there are no points – instead, a large person standing a bridge over the tracks. **Do you push Fatty into the path of the train to save the five?**

I chose yes in both scenarios.

> "The needs of the many outweigh the needs of the few." – Spock

Interstingly, although I was with the majority in the first scenario, I was in the minority in the second. I don't understand this difference, except for the difference between operating a control like the lever, and physically acting on another human being – the former seems less connected to the action, perhaps.

Other scenarios How does the trolley problem change when the 5 are convicted rapists? Children? When Fatty is a scientist working on a cure for Leukemia? When the 5 are Mountain People (insert your own other-class of person)? Tories? Welsh? Rednecks?

Making moral decisions

So, are ethics about choices, or outcomes? Rules or consequences?

MIT are trying to build a picture of moral acceptability with their crowdsourcing moral machine project in which you are presented with multiple scenarios (like driverless car choices) in which you decide the path the vehicle should take, usually resulting in somebody or something dying. It is easy to become quickly abstracted from the awfulness of the choices you make in this game.

Guidelines in ethical frameworks for machine choices are broadly divided into four categories:

- Do good
- Minimise harm
- Respect human autonomy
- Be just or fair

Someone should tell Peugeot about the third one. My wife's 2008 has a really irritating habit of grabbing the wheel if you change lanes without indicating, thinking that you've fallen asleep. I haven't found a way of switching that off yet, but it is really disconcerting, especially on a long trip at night with no other vehicles around.

Transparency of the algorithm and accountability are increasingly being emphasised. These frameworks have become regarded as inadequate as they offer a way for corporations to hide behind them in what is called, "ethics washing". This is a problem for all rules or specifications, or checklists.

Decision-making in Scottish Teacher Education

In my own application, teacher education[1], the GTCS Professional Standards are meant to provide an objective benchmark that describes the competencies and skills of all teachers in Scotland. We know that the application of these standards is in the hands of the profession itself and therefore wildly variable in their interpretation. The standards themselves are written in ambiguous and wishy-washy language, like most things in state-provisioned education, and so are highly subjective and open to – interpretation or abuse, depending on your view of an individual situation.

> "*Universal law is for lackeys. Context is for kings.*" – Captain Lorca, StarTrek Discovery

[1] We don't like "teacher training" because we like to think that teaching is a profession, like the law, or the military. It isn't, of course, but we sustain the pretence for ourselves, even if nobody else in society beleives it.

Moral decision-making applied to data ethics

A couple of questions are asked in a section making the links between moral decision making and data ethics. The first, "*Should we require people to give their DNA to a gene bank?*", screams at me, perhaps because of my age and the closeness people of my age have, although not by first-hand experience, of the horrors of the Second World War. Everything about centralised reporting of ethnicity makes me recoil: I never provide information about my ethnicity, and I would need a very good reason before I ever did. We are forgetting: something that gets me in trouble almost every year, with very real regret and a deepening sense of injustice[2].

The second question relates to social media and their handling of "hate speech". This is topical: I have just deleted my personal Facebook account[3] after they, once again, dismissed my objections to posts intended to whip up hatred of muslims, or other racial or ethnic groupings. In every case I have raised, the posts have not been found to breach codes of conduct, illustrating what is described as ethical washing (see above). To my cost – I have cut off friends and family with whom I am only connected this way – I have deleted my account.

A thought experiment

We are tasked with making our own thought experiment to allows us to examine our beliefs and "surface" factors that influence our judgement. This is mine.

The question Is it OK to kill to save lives in a war when you are not a combatant?

Parameters In the context of war, it is given that the combatants of one side are allowed to kill the combatants of the other. What constitutes "combatant", however? Is it anyone who wears the uniform, or stands behind the barricades? What about observers?

The story During an African war, a soldier of an impartial peacekeeping force (soldier A) is invited to take a ride in a helicopter on a routine supply drop to a station in the bush. Soldier A has not been explicitly told that he is not allowed to ride in the vehicles of either side. The route is a well-known safe corridor, well policed by peacekeepers and resepcted by both sides in the war. Soldier A takes the ride, seated in the open side door of the helicopter next to a mounted cannon. En route, the helicopter comes under fire unexpectedly. The pilot takes evasive action by banking sharply to

[2]When I call people out for their disrepect of the remembrance observation, it is always me that is critiqued for not being kind, or collegiate. People can be so ungrateful and selfish. The injustice of such treatment makes me wonder.

[3]I am keeping, uncomfortably, a couple of social media accounts going that are related to my media and technical interests. Incidentally, I also deleted my personal Twitter account, but not for this reason. I'm generally a bit fed up with the whole Internet at the moment.

right. Soldier A can see smoke coming from a group of boulders on the ground directly in front of him. He realises that his life, and the lives of the others in the helicopter are in danger. He hears the pilot on the intercom shouting, "Shoot them! Shoot them!"

You are soldier A. Do you use the cannon and try to save yourself and the crew?

Legal and ethical considerations

Legal

Well, let's not pretend that there's much of a connection between law and morality. Laws are instruments of power. They belong to the powerful. Even at mate's rates, a lawyer's fees are vulgar and shameless. Even if it were accessible by the common people, legal decisions have always been historically hysterical, wildly inconsistent and woefully inadequate. It's noticeable that the great democratisers of the law have been people like Napolen and Hitler. It's about time we replaced the whole system with a machine.

> "*The first thing we do, let's kill all the lawyers.*" – Henry VI, Part 2, Act IV, Scene 2 (1591)

The laws of the land are the first and greatest example of what the course describes as "ethics washing", the practice of corporations who create "codes" to *lure the public into a false sense of security and earn undeserved trust.* Laws make us feel safe but do nothing to make us safe as enforcement and monitoring are lacking.

In the same way that the discussion in the course speaks of ethics guidelines being used by companies "to ward off regulation by the state through formal laws", so government uses laws to ward off revolution and disorder.

There is a tension in data ethics between the need and cost of self-regulatory codes of conduct, and external regulation through the courts. The course summarises data ethics and the law:

Hard ethics is needed to understand and interpret laws, and to make sure legislation is followed.

Soft ethics are normative rules that tell us how to behave morally, whether the law addresses our actions or context.

Compliance with the law is normally necessary for ethically correct conduct, but may not be enough for it.

Ethical? Nothing ethical going on here.

I skimmed the transcript of the role-play presented by the course (I am out of week already) and also the discussion on the ethics of setting cookies for the participants of ths MOOC. Much angst, I see.

Tracking is an issue for me and therefore I take steps to mitigate it. I don't allow setting of third-party cookies, and have a policy of clearing all cookies on exit from a browser session. I also use different browsers. Tracking is difficult under those circumstances: this is working for me. How do I know? When I get advertising, it's usually totally irrelevant, which gives me some comfort, but I'm not complacent about it. I hate advertising: it's one of the manifestations of what is truly awful about human beings.

Solutions like the GDPR persmissions dialogues are no better than the advertising itself. They doesn't solve any problems, they just gets in the way and decrease the chances of me using the sites that push them into my face. I make use of ad blockers, brutal cookie policy, and text readers when using the web.

Interestingly, the link to the EC directive yields a 404 (not found) error. The EdX FAQ is a good example of the dishonest corporate response to such issues: it presents questions "frequently asked" by users which it doesn't answer.

One of the difficulties of this dominance of the common habit of making a website, for whatever reason, of the legal constraints is that the people that benefit most from this are lawyers. They get fat on the friction of common activities. I have been making websites since the early 1980s before anyone knew what the WorldWide Web was. Most of them are *pro bono* educational sites, or services for the common good. Some have been for profit but not in the "YouTuber" sense: I have charged fees to compensate for the time it has taken. That time has not included exhaustive compliance checking with the various laws and expectations. I have always worked on the basis that I haven't got any money, so am not worth pursuing in litigation. So far, so good.

At no point has it been made clear why cookies are necessary. Disabling cookies on the EdX site results in denial of access immediately on refreshing the page. So, the first purpose is to grant access to the site (which can be done in other ways easily enough, and more securely).

Information, control and power

Formalism Legal choices based on logic only: what you did, and what the law is.

Realism Your conviction depends to a large extent on what the judge ate for breakfast.

A task is given in which students are asked to decide if building an app to influence proceedings based upon what is known from an AI analysis of jury members, is ethical. I picked the red envelope, but both answers are characterised as "incorrect" for their impact on justice. What is not presented is a choice to undo the injustice of the system itself, in which jury manipulation is permissible.

Common ethical issues

The rapid increase in the availablility of data, particularly on consumers and their habits, has led to the rapid increase in its manipulation and exploitation, leading in turn to ethical challenges as community dependence on the systems that gather the data has produced apathy, indifference, or a sense of powerlessness in resisting it.

The very term *algorithm* has been re-purposed by social scientists to further add to the problem by obfuscating the issues as they attempt to take ownership of the matter.

The transparency of algorithmic actions has been problematised. A new fear is whipped up about how we are being manipulated – we are, of course, and it's our fault because we are weak, stupid or lazy. The fear of processes that are not simply understandable exacerbates the feeling of powerlessness. Explainable Artificial Intelligence, or **XAI** is a recent field that tries to push back against this fear. Why? Because that fear results in the populus avoiding participation in the behaviours that enslave them.

> "You do look glum! What you need is a gramme of soma." (Huxley, 1955).

The issue of bias in algorithms seems to be poorly understood, perhaps because of the belief of the academics that they hold some kind of moral authority to impose blind egalitarianism, or a kind a false neutrality on systems that appear to present biases. This, from the course:

> This is because they are designed by humans and trained on data generated from and by us, and therefore hold the potential to encode discrimination within decisions and predictions.

What seems to be suppressed is the possibility that the data may lead to unpalatable conclusions. There are more blacks per capita in jail because the system is racist. Other conclusions are possible, just not acceptable, and we seem to be able to bend ourselves into all kinds of shapes to avoid them.

Consent

Consent is presented as a way to cleanse the abuse of data and hold harmless those that use it. This is difficult, of course and those who operate systems in my

experience are often ignorant of consent requirements (in the use of submitted assignments to inform future students, for example), or simply ignore them in the hope that *what the eye doesn't see, the heart doesn't grieve over*. Consent doesn't work.

Week 2: Crime and Justice

- Ambiguous Ethical Issues
- Crime, Justice and Technology
- Bias
- De-biasing Algorithms
- Fairness
- Data Justice

Ambiguous Ethical Issues

Predictive policing

Another talking head, this time revealing the extent of the use of technology in policing in the UK, as reported by the leftist lobby group NCCL[4](Couchman, 2019). Here, the media response to the use of AI to predict things like potential hot-spots for new crime, or even potential criminal behaviour of individual citizens, is noted. They made links to the Hollywood Film, "Minority Report" which was based on a similar theme.

The fear is that existing (negative) biases will be amplified: this is ceratinly a resonable fear and hinges on the data that is used to train the machines, which must be historical data.

> "..for I the Lord thy God am a jealous God, visiting the iniquity of the fathers upon the children unto the third and fourth generation..."
> – Exodus 20:5

What the video fails to do is to identify this report's finding that predictive policing is not in use (and has not been used) in Scotland or Northern Ireland. I found that odd, as the presenter described the report as being "*about the rise of predictive policing across the UK.*"

[4]AKA "Liberty", led from 2003 – 2016 by Shami Chakrabarti, who was given a peerage by Jeremy Corbyn.

The attraction for police forces is clear: greater efficiency in the deployment of sparse resources is a desirable feature of any publicly-funded service. Similarly, benefit and social work agencies have the same gains to make, and the different application of deciding fair and consistent sentencing in the criminal courts is also clearly desirable *if it is sound.* That requires awareness of the full feature set of the data used to train the systems and the ability to compensate for errors by the application of appropriate weightings or bias.

Algorithms and mathematical models

This section provides example cases of "algorithmic bias" in welfare services that are in part computerised. Interestingly, it doesn't mention China's Social Credit System which "rates" individuals according to trustworthiness: participation in the system became mandatory this year for Chinese citizens. Social Credit is economic and social reputation of individuals and business entities. Reputation is earned and lost by behaviours: good acts like giving blood or doing voluntary work are positive, bad acts like jaywalking, using your sister's transit ID card, jumping a red light, are negative. Credit determines how accessible services and rights are to you: university places, hospital procedures, employment, and so on. Untrustworthy citizens are posted on social media channels and posters.

Automating poverty task

Time is very short this week, so I skimmed the articles but did not participate in the crowdsourcing discussions[5].

Crime, Justice and Technology

Ferguson (2017) describes how policing moved from a *clinical* to an *actuarial* model: from expert, individual decisions unconstrained by the parameters of a pre-designed model, to fitting people and their behaviours into categories derived from historical anaylsis. This is clearly a fundamental data mistake: to close off a model to new experience and insight is an accountant's blunder[6].

[5] I find always these a chore, and of little value in online courses like this. It's become a formula for the MOOC: I enjoyed these once, and even gone on to connect in real life with people I've met in the discussion spaces. However, in the spaces where this has been trotted out as some kind of lazy self-service pedagogical device, it always falls flat as an empty task. I get nothing from these, and almost never get commentary on my contributions, including this course so far, in which there are a number of people talking in threads, but not to each other. Pearls before swine, darling.

[6] One of my students told me today that she had started reading *Foundation* (Asimov, 1991), which I still revere as the paradigm of modelling to which we ought to aspire. It remains as far from our clumsy, incompetent implementation of technological advances as one can imagine.

Justification for the use of such tools is offered within the terms of a social contract, in which citizens surrender certain freedoms (e.g., not to kill each other – obviously a freedom the NRA goads the rest of us to prize from its cold, dead, hands) in exchange for the benefit of protection against others who might harm us.

Two types of predictive policing are identified in the course:

- Predictive mapping
- Individual risk assessment programmes

Predictive mapping

This relates data on the time and place of crimes, perhaps also with additional data on the type of crime (but not the criminal) and uses this to make a prediction or forecast of likely "hot-spots" to which police resources may be deployed in an attempt to mitigate. There are a number of reports cited for further reading but the report from Chile (Contreras, 2019) is not atypical of much media coverage of the use of this kind of technology. Here's how it opens:

> El 29 de agosto de 1997, a las 2:14 AM Skynet toma conciencia de sí misma. Skynet es una inteligencia artificial que lidera el ejército de las máquinas que quieren exterminar a los humanos pues los considera una amenaza para su propia supervivencia. Skynet está basada en una red neuronal que funciona en la nube y que maneja todos los aviones y armas no tripuladas de los Estados Unidos de Norteamérica. Para eliminar a los seres humanos desata una guerra nuclear y el posterior apocalipsis.
>
> On August 29, 1997, at 2:14 AM Skynet becomes aware of itself. Skynet is an artificial intelligence that leads the army of machines that want to exterminate humans because it considers them a threat to its own survival. Skynet is based on a neural network that operates in the cloud and handles all the planes and unmanned weapons in the United States of America. To eliminate humans it triggers a nuclear war and subsequent apocalypse.
>
> – translation by DeepL

Individual risk assessment programmes

Worse than that, of course, is where it gets personal: connecting *individual* with *risk assessment* is a hot potato in education, let alone policing. Perhaps because it starts with the stance that an individual person presents a risk to "us", making them implicitly "other".

One of the difficulties with predictive policing is that it undermines one of the purposes of criminal law, rehabilitation, because it targets previous offenders for closer police attention in their community. This must also significantly impact on deterrence: if the police are focusing on the ex-cons, they are less likely to be looking at the rest of us, *increasing* the likelihood that we will be deterred from criminal activity[7].

The LAPD stopped using its LASER tool in 2019 but continues to engage in data-driven policing.

Sentencing and parole

The use of data is not a new feature of sentencing and parole decisions world wide and these uses have led to sustained bias scoring (and thereby sentencing or incarceration decisions).

Bias

The course resorts to Webster's, of all places, to find a definition of bias. My background understands the term as:

> "A steady voltage or current applied to an electronic device" – (*bias, adj., n., and adv. : Oxford English Dictionary*, n.d.)

For me, bias and its sister, discrimination are not inherently Bad Things. The ordering of the meanings of the word in these dictionaries offers perhaps a cultural explanation for a bias of the course writers, who seem to be seeking to associate injustice with this word. The OED's principal meaning of *bias* is *slanting, oblique* (*bias, adj., n., and adv. : Oxford English Dictionary*, n.d.).

Algorithmic bias

One problem with the current criticism of "algorithmic bias" is that the scapegoat is either the algorithm itself, or the programmers that make it. It seems unfashionable to respond, "fair comment" when outcomes of data analysis go against our modern, fragile, sensibilities:

> "Google's ad-serving system showed an ad for high-paying jobs to men much more often than it did for women" – from CMU research, cited in Kirkpatrick (2016)

[7] I stopped watching the video sequences included in the course because of the intrusive advertising – even interrupting a clip on the use of data by police in the USA to show me a 5-second Condé Nast advert. I was shown the same ad 5 times in one video clip. Am I naïve to expect that education really ought to be free of such exploitation?

Clearly the targeting of the ad wasn't based upon the single variable of gender: other factors are significant in the selection *and are ignored in the reporting* to make a more sensational commentary. This is the real difficulty for me here: that we ignore the manipulation of our responses by irresponsible and lucrative[8] articles like these.

Big Data and implicit bias

A reading from this week (Barocas, 2014) also examines the meaning of words when suggesting ways of tackling the problem of implicit bias in algorithms derived from historical data. I like the moderate language used in this essay:

> "If data miners are not careful, the process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination." – Barocas (2014), p. 673

> "Discrimination may be an artifact of the data mining process itself, rather than a result of programmers assigning certain factors inappropriate weight" – ibid. p. 674

So, we can step away from the tribal frowning at the technology and the people who program it and recognise that our own past behaviours have created this learned behaviour, in the same way that it is created in our own attitudes and biases. Barocas (2014) suggests this can be done via the lens of (US) Title VII (*Civil Rights Act of 1964 - CRA - Title VII - Equal Employment Opportunities - 42 US Code Chapter 21*, 1964) and by doing what I consider to be the obvious, which is to understand how these algorithms get their biases in the first place:

> "Data mining takes the existing state of the world as a given and ranks candidates according to their predicted attributes in *that* world." – Barocas (2014), p. 731, orginal emphasis

De-biasing Algorithms and fairness

Fairness-aware machine learning is a term used in an EU report (Tolan, 2018) that asserts that *fairness* in this sense depends on the domain in which the model is being made and therefore the fairness constraint applied have to be specific to that domain.

The group fairness approaches in the EU report apply political or popular biases like:

[8] We like to be titillated with stories like this: we buy newspapers for them, we watch the channels that serve them, and subscribe to media that feeds them to us.

> "the share of defendants classified as high risk should be equal across different protected groups" – Tolan (2018), p8

This is called *demographic parity* and tries to neutralise an aspect of the source data: suppressing one truth in the name of another, perhaps. In another, *calibration* is applied such that "the proportion of people re-offending is the same across protected groups" (ibid, p.10), and to achive *similar people should be treated similarly* is considered a non-trivial task on account of deciding on what data is required to identify how similar two people are.

"Fairness through unawareness" is one method described by which algorithms might be made to mitigate bias: this seems to take us full circle back to where the problem began, in which the use of historical data is used to train the model from the prior behaviours which themselves have included hidden biases.

Data Justice

The framework called Data Justice looks at both social and technical aspects of machine bias. The idea is to oppose exacerbating social injustice by the rapid adoption of technologies that embed data which itself may include past social injustice. The course points us again at Couchman (2019) for its stance on "policing by machine" but also introduces us to one writer's proposal for redressing the power balance back in favour of the citizen, namely (in)visibility, (dis)engagement with technology and antidiscrimination (Taylor, 2017).

The first of these calls for greater transparency of what and how data is collected and used, and allows individuals to choose not to be part of it. The second relates to the latter point and calls for greater control for the individual on how (or whether) they participate in the data markets. Finally, individuals should have the right to call out bias or unfair treatment at the hands of data-informed systems. This, I think is particularly important for public services.

Increasingly, watchdog groups are being established around the world to raise awareness and facilitate action against the abuse of data. The UK's Data Justice Lab sits within Cardiff University's media school. Interstingly, the FAT/ML[9] website hasn't been updated for the past two years, which might suggest that nothing much is happening in the group.

[9]Fairness, Accountability and Transparency in Machine Learning, led by Solon Barocas, of Barocas (2014) and Microsoft.

Week 3: Home and City

- The Internet of Things
- The Smart Home
- The Smart City
- Design solutions

References

Asimov, I. (1991). *Foundation*. Bantam Books.

Bardi, A., Holloway, R., Lönnqvist, J.-E., Bevington, P., Hanel, P. H. P., Maio, G. R., Soares, A. K. S., Vione, K. C., De Holanda Coelho, G. L., Gouveia, V. V., Patil, A. C., Kamble, S. V., & Manstead, A. S. R. (2018). *Cross-Cultural Differences and Similarities in Human Value Instantiation*. https://doi.org/10.3389/fpsyg.2018.00849

Barocas, S. (2014). Big Data's Disparate Impact. *California Law Review*, *104*(671), 671–732.

bias, adj., n., and adv. : Oxford English Dictionary. (n.d.). Retrieved November 16, 2020, from https://www.oed.com/view/Entry/18564?result=1&rskey=pOoRi9&

Civil Rights Act of 1964 - CRA - Title VII - Equal Employment Opportunities - 42 US Code Chapter 21. (1964). https://finduslaw.com/civil-rights-act-1964-cra-title-vii-equal-employment-opportunities-42-us-code-chapter-21

Contreras, D. V. (2019). *Batallas 3.0: Inteligencia Artificial y algoritmos versus delincuencia en Chile*. https://www.theclinic.cl/2019/07/25/batallas-3-0-inteligencia-artificial-y-algoritmos-versus-delincuencia-en-chile/

Couchman, H. (2019). *Policing by Machine* (pp. 1–48). Liberty.

Ferguson, A. G. (2017). Predictive policing. *Washington University Law Review*, *94*(5), 1109–1189.

Huxley, A. (1955). *Brave new world : a novel*. Penguin Books in association with Chatto & Windus.

Kirkpatrick, K. (2016). Battling algorithmic bias. *Communications of the ACM*, *59*(10), 16–17. https://doi.org/10.1145/2983270

Shakespeare, W. (1591). *Henry VI, part 2: Entire Play*. http://shakespeare.mit.edu/2henryvi/full.html

Taylor, L. (2017). What is data justice? The case for connecting digital rights and freedoms globally. *Big Data and Society*, *4*(2), 1–14. https:

//doi.org/10.1177/2053951717736335

Tolan, S. (2018). *JRC Digital Economy Working Paper 2018-10 Fair and Unbiased Algorithmic Decision Making : Current State and Future Challenges. December*.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Wagner, B. (2019). Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping? In M. Hildebrandt (Ed.), *Being profiled* (pp. 84–89). https://doi.org/10.1515/9789048550180-016