The Philosophy and Ethics of Artificial Intelligence

Dr. Sanjay Modgil

Overview of Course

- Introduction
- Intelligence
- Consciousness
- Reasoning and Communication
- Ethics and Morality
- Algorithms
- For good or bad ? AI in Medicine, AI in War
- Superintelligence and the Value Loading Problem
- AI and Human Society
- Review/Revision

Algorithms

- Accountability and Transparency
- Liability and Responsibility
- Algorithmic Bias

Accountability and Transparency

- How do we ensure *accountability* when algorithms make decisions: how and to whom or what, do we assign responsibility / blame when things go wrong ?
- Many important decisions historically made by people are now made by computers. Algorithms count votes, approve loan and credit card applications, target citizens or neighbourhoods for police scrutiny, select taxpayers for auditing, grant or deny immigration visas ….
- Tools available to policymakers, legislators, and courts, developed to oversee human decision-makers, often fail when applied to computers. For example, how do you judge the **intent** of a piece of software?
- Because automated decision systems can return potentially incorrect, unjustified, or unfair results, additional approaches are needed to make such systems accountable and governable – especially when decisions have life/ death consequences

Accountability and Transparency

- So how can we ensure accountability ? That is, how can one demonstrate to the public at large and to regulatory and legal bodies, that automated decisions comply with key standards of legal fairness and ethical fairness ?

Transparency - the ability to examine and explain how algorithms make decisions which is important in ensuring accountability.

Transparency not only allows one to *demonstrate* ethical fairness but also provides *incentives* to ensure compliance with ethical standards

In Plato's *Republic*, Protagoras considered the Ring of Gyges which makes wearer invisible. Protagoras argues that wearer would commit all manner of wrong doing.

Fears that without our knowledge we may be manipulated by powerful machines or very powerful corporations making use of opaque (non-transparent) AIs = modern take on Ring of Gyges myth. But transparency and hence accountability will lessen likelihood of this.

Accountability and Transparency

However:

- Explaining how Machine Learning algorithms make decisions is a major unsolved issue – ***the black box problem*** – even for AI researchers
- Sometimes transparency is not desirable. E.g. when:

- explaining a decision may reveal private data about an individual (so violating EU General Data Protection Regulation (GDPR))
- transparency may be abused, e.g. when deciding which tax returns to audit, or whom to pull aside for security screening at airport, then may need to be partly opaque to prevent tax cheats or terrorists from gaming the system.

Arguably demanding transparency involves weighing up the harms and benefits – we wouldn't want transparency to be a universal value/virtue or deontologically insist on transparency

Procedural Regularity

When AI algorithms used in narrow and predictable contexts, a basic requirement is *procedural regularity*: each participant will know that the same procedure was applied to her and that the procedure was not designed in a way that disadvantages her specifically.

- **Specifically tools for procedural regularity can assure that**

- The same policy or rule was used to render each decision.
- The decision policy was fully specified (and this choice of policy was recorded reliably) before the particular decision subjects were known, reducing the ability to design the process to disadvantage a particular individual.
- Each decision is reproducible from the specified decision policy and the inputs for that decision.
- If a decision requires any randomly chosen inputs, those inputs are beyond the control of any interested party.

Procedural Regularity

- **Techniques for ensuring procedural regularity**

- **Software verification** which uses logic based techniques to formally prove that program's behaviour satisfies certain desirable properties = invariants, or facts about a program's behaviour that are always true regardless of a program's internal state or the input data

  For example, *model checking* a program - exhaustively testing for all possible inputs to ensure that an invariant is never violated

- **Cryptographic commitments** can be used to ensure that the same decision policy was used for each of many decisions (especially useful when full transparency is not desirable)

(for more details see: https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=1888&context=faculty_scholarship)

ACM Principles for Algorithmic Transparency and Accountability

(A)ssociation for (C)omputing (M)achinery is the world's largest educational and scientific computing society

Awareness Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use, and the potential harm that biases can cause to individuals and society.

Access and Redress Regulators should encourage the adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions.

Accountability Institutions should be held responsible for decisions made by the algorithms that they use, *even if it is not feasible to explain in detail how the algorithms produce their results*

Explanation Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts.

ACM Principles for Algorithmic Transparency and Accountability

Data Provenance documents the inputs, entities, systems, and processes that influence data of interest (e.g. training data) and so provides a historical record of the data's origins, what processing it underwent (e.g. by providing links to the algorithms it was processed by, together with the input parameters), and the systems/entities that processed the data.

Such a record should be maintained and ideally accompanied by an exploration of the potential biases introduced by the human or algorithmic data-gathering process.

Public scrutiny of the data provides maximum opportunity for corrections.

However, concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to abuse the system can justify restricting access to qualified and authorized individuals.

.

ACM Principles for Algorithmic Transparency and Accountability

Auditability Models, algorithms, data, and decisions should be recorded so that they can be audited/inspected in cases where harm is suspected.

Validation and Testing Institutions should use rigorous methods to validate their models, and document those methods and results.

In particular, they should routinely perform tests to assess and determine whether the model generates discriminatory harm. Institutions are encouraged to make results of such tests public.

AI, Liability and Responsibility

Accountability Institutions should be held responsible for decisions made by the algorithms that they use, *even if it is not feasible to explain in detail how the algorithms produce their results*

But with increasingly autonomous AI systems who or what is responsible and hence liable when things go wrong ?

Driver error causes 94% of conventional (non-autonomous) car crashes.

Fully and partially autonomous vehicles (AVs) could substantially improve that number.

Although there have been over 30 accidents involving AVs in the state of California since 2014. So crashes, injuries, and fatalities will not disappear - AVs will crash into one another, into conventional cars, pedestrians and bicyclists.

When it comes to accidents involving AVs, who is responsible?

AI, Liability and Responsibility

Current assumption is that the companies behind the software and hardware are legally liable, not the car owner or the person's insurance company

But the line between human and machine liability isn't always clear

To date, several accidents show that AV was warning driver to disengage autopilot mode and take control of the vehicle. Is human driver then responsible ?

Eventually, AVs may not have a human at the wheel or even have a steering wheel, so how could a human passenger intervene? Carmakers will have to take responsibility.

Note that in the one and only recorded case of a fatal AV accident in Arizona in March 2018, the initial police investigation indicated that pedestrian at fault.

But what do current frameworks for legal liability say ?

AI, Liability and Responsibility

Could an AI system be held ***criminally liable***, which usually requires an action and mental intent ? Three scenarios:

1) ***perpetrator via another***: applies when an offense has been committed by a mentally deficient person or animal, who is therefore considered innocent. But anybody who instructed the mentally deficient person/animal can be held criminally liable. Eg a dog owner who instructed the animal to attack another individual.

An AI program could be held to be an innocent agent, with either the software programmer or the user being held to be the *perpetrator-via-another*

2**) *natural probable consequence***: Eg in 1981 an AI robot in a Japanese motorcycle factory killed a human worker. The robot wrongly identified the employee as a threat to its mission, and calculated that the most efficient way to eliminate this threat was by pushing him into an adjacent operating machine, killing him instantly, and then resuming its duties.

The key question was whether programmer knew this outcome was a probable consequence of its use.

AI, Liability and Responsibility

3) ***Direct liability***: which requires both action and intent. Action is straightforward to prove, but intent ? We return to this issue next, when we discuss the notion of **moral responsibility** and punishment for direct liability

But criminal liability may not apply → the matter would have to be settled with civil law. Then a crucial question will be whether an AI system is a *service* or a *product*

If it is a product, then product design legislation would apply, based on a warranty for example.

If it is a service, then the rules of negligence apply. In this case, the plaintiff (the person who brings the case to court) would have to demonstrate three elements to prove negligence:

1) The defendant had a (deontological) duty of care
2) The defendant breached (violated) that duty.
3) The breach caused an injury to the plaintiff.

We revisit some of these issues when we discuss the ethics of AI in medicine. However, things get really complicated when future AI becomes more human-like (AGI) and perhaps even superhuman. Could such systems be criminally liable ?

Moral Responsibility, Free Will and Justice

In philosophy and law, the issue of whether an agent can be said to be ***morally responsible*** is crucial in deciding whether the agent deserves praise, blame, reward or punishment for an action

Typically, a person is morally responsible for a crime if they:

1) were conscious (they 'understood') that the crime would cause harm (they understood the ethical impact)

2) were free to do otherwise –they exercised their **free will** in choosing to commit the crime rather than not commit the crime. Assumption of free will goes hand in hand with the idea that there is a **self** – the self that freely chooses.

The self: the I that I am – the you that you are – the sense one has that they are more just a physical body, but that there is a self that is you, that thinks, that is the *subject* of conscious experience (suffering, joy, pleasure …) an identifiable self that exists and persists over time. The self that exercises free will (freely chooses amongst options).

Moral Responsibility and Full Ethical Agents

Note conditions for moral responsibility:

- **conscious understanding of ethical impact and**
- **freedom to choose otherwise**

also characterises Moor's notion of a full ethical agent

Moral Responsibility and Justice

Moral responsibility means that the law can exact a penalty on the person (imprisonment, fines, etc) in order to

i) *protect* society from the criminal, and protect the criminal from society;
ii) *deter* others from committing the crime;
iii) *rehabilitate* the criminal so that (s)he doesn't commit the crime again;
iv) punish the person (*retribution*) – which might be understood as society seeking vengeance

Retribution is effectively punishing the **self** for having **freely** chosen to commit the crime. Notice that deciding the right kind of penalty on the basis of achieving i), ii) and iii) can be understood as deciding the right kind of penalty on ***consequentialist*** grounds. Why ? What would be the consequences of retribution that are not covered by i), ii) and iii) ?

Could an AGI or SuperAI be morally responsible for a crime ?

We have already suggested that it is an open question as to whether AI machines can be conscious and so consciously understand the ethical impact of their actions

AI Selves and Free Will

But would there be a distinct AI self that freely chooses amongst options ?

Isn't it the case that there is just algorithms that process information about the world, the effects of actions, and then makes choices (selects actions) that maximise utility and/or follow a set of rules and/or choose actions on the basis of rewards ?

Even the decision as to how one judges/evaluates the utility of states of affairs is an essentially algorithmic decision based on given information.

Who is the self in the AI (the 'ghost in the machine') ?

Conclusion: **There is no self, there is no free will, on the basis of which we can say that an AI has moral responsibility**

AI Selves and Free Will

There is no distinct AI self that freely chooses amongst options

So perhaps a consequentialist framework can be used to decide penalties in order to ensure:

i) *protection* of society from the AI;
ii) *deter* others AIs from committing the same crime (perhaps by penalising developers even though they are not "morally responsible" ?)
iii) *rehabilitate* the AI so that it doesn't commit the crime again (i.e., change its code);

the penalties need to ensure that the action selection modules of other AIs (*deterrence*) and the 'criminal AI' (*rehabilitation*) are appropriately modified so that the criminal action is not chosen;

iv) But retributive punishment wouldn't make sense – there is no self that freely chooses and that therefore deserves punishment (*retribution*)

Human Selves and Free Will

But doesn't the same reasoning equally apply to a human ? Is there a distinct human self that freely chooses amongst options ?

It is common-place to think of the self as the "chief executive" overseeing the workings of the mind – the I (the ego/ the self) that is the subject of experiences, that has thoughts, that freely chooses etc etc

Here are some other things we might think about our *selves*

- I was once very small
- If there are no major accidents, I will become old
- My body may change but I will remain the same
- I might have been born at another time or place
- It would be great if I could be Leonardo Dicaprio for a day
- My self or soul will survive after my body has died, to then ascend to heaven (Christianity, Islam) or be born into another body (Hinduism)

It is common-place to also therefore think of a self ('the soul') that is separate from the body, and that some believe survives in some form, after death

The Human Self

But when we closely examine the idea of a self, all there is is just neurons firing, giving rise to thoughts and experiences.

Where is the distinct *inner* self that has these thoughts and experiences ? The self that exercises free will (freely chooses) ?

And if there is a distinct self, inside your head, having thoughts and experiences, how would you explain how this self inside your head has thoughts/experiences ?

The Self is an Illusion

Modern neuroscience, philosophy of mind (and Buddhism !) all converge on the idea that the notion of a distinct self - someone looking out at a world out there, and also watching our own thoughts pass by – is an illusion

There are only mental processes. There are streams of thoughts, sensations and perceptions passing through our brains, but there is no central place where all of these phenomena are organised, no CEO/self that weaves together these streams to yield a unified integrated experience

> When Descartes famously said “I think therefore I am” what he should have said is: “Thinking is being”

The sense of a self is how we consciously experience the integration of information

For evolutionary reasons it makes sense that we evolved this experience of a distinct self that is so crucial to our sense of identity and who we are.
Because **self** preservation implies **body** preservation which in turn ensures that your genes will passed down to the next generation

Free Will is also an Illusion !

The laws of physics tells us that everything is determined by what has gone before. It's all just chains of cause and effect (also "law of interdependent origination" in Buddhist philosophy).

Given a choice between A and B, the choice we make is determined by neuronal processes in the brain, which in turn obey the laws of physics.

At a higher *level of explanation*, the choice is fully determined by our genes and our environment (the sum total of our experiences prior to making the choice)

Levels of explanation: One can describe the working of a car engine in terms of the physics of all the particles that define the physical structure of the engine. But a higher level of explanation is more meaningful and informative (the fuel burns and so releases an expanding gas that forces pistons to move ….)

Free Will is also an Illusion !

Note**, we do have choices** – but our choices are decided by all that has happened before (our environmental history), and our genes, and ultimately explainable in terms of the laws of physics in which every event that takes place in the world (including choices we make) is determined by a preceding chain of cause and effect.

The conscious experience of making a choice – *what it feels like when make a choice* - is the *feeling* that we freely make the choice.

Conclusion: Free will, as commonly understood, as being “really free” can only mean a choice that is not at some level determined by the laws of physics. Otherwise it would not be commonly understood as free will.

Free will is a concept that is only meaningful when we refer to what is feels like (the nature of our conscious experience) when we make a choice.

Justice for Humans

Retributive justice is about exacting revenge against the self that freely made a choice, but if there is no self freely making a choice, then we should abandon the idea of retribution, and focus on the other reasons for penalising criminals

i) *protect* society from the criminal, and protect the criminal from society; ii) *deter* others from committing the crime; iii) *rehabilitate* the criminal so that he doesn't commit the crime again; ~~iv) punish the person (*retribution*)~~

These reasons are essentially consequentialist (ultimately utilitarian): Penalties are imposed to maximize well being. Of course, we treat people **as if** they had selves and free will and so are morally responsible, just as we think of ourselves as having free will and being morally responsible. By treating people this way we feel justified in penalizing them. The point is that the nature of the penalties should be decided on consequentialist/utilitarian grounds

Penalties act on the choice function of the criminal (rehabilitation) and others (deterence), to ensure the right choices are made in the future

A caveat: for more serious crimes, it may be that if a criminal is not punished, society would break down because a lack of faith and trust in the justice system. So there may be utilitarian arguments for having some retributive component !

What does this mean for AI systems of the future

So to sum up what does this all mean for the AI systems of the future

- They, like us, will **not** have distinct *real* selves or *real* free will
- The issue of morally responsibility is irrelevant to deciding how we should react when things go wrong and we/they make wrong choices.
- ➔adopt a consequentialist/utilitarian approach – the penalties should be such that we ensure:
- The AI in question should be taken out of use (society is *protect*ed from the AI) and/or rehabilitated so that it doesn't commit the crime again, e.g. by changing the decision making algorithm) and penalties imposed (on the companies, corporations and institutions) so as to *deter* development and use of AIs that would commit similar such "crimes"

Note that it is likely that we will treat these future AIs ***as if*** they had free will and were morally responsible. It may then be that if an AI commits a serious crime, the public's faith and trust in AI will only be preserved if the AI is 'punished' - 'decommissioned' i.e., turned off (assuming that they do not suffer ?)

What does this mean for AI systems of the future

Finally, recall that in Moor's categorisation of Ethical Agents

Explicit Ethical Agents: Machines that can reason what is the best action in ethical dilemmas, and novel situations that may not have been anticipated by designers, using ethical principles – explicit representations of ethics e.g. in the form of deontic rules.

Full Ethical Agents: Machines that can be said to have 'moral agency' and are able to justify their moral judgements. In philosophy and law moral agency equates with *moral responsibility* – if one understands the ethical impacts of one's actions and freely chooses action that is ethically wrong, then one is morally responsible in the eyes of the law, and so should be held accountable for one's actions (i.e., punished)

Suppose we agree that in principle, it is not possible to definitively know (know with 100% certainty) whether an AI is conscious and so are conscious of the fact (i.e., 'understand') that the crime would cause harm. Given what we have said about free will not being meaningful when it comes to AI (and indeed humans) does it make sense to then distinguish full ethical agents ?

Algorithmic Bias

Bias in Algorithmic Systems describes systematic and repeatable errors in a computer system that creates **unfair** outcomes, such as favoring one arbitrary group of users over others.

Algorithmic systems in this context refers to the combination of algorithms, data and the output deployment process that together determine the outcomes that affect end users.

For example, a credit score algorithm may deny a loan without being unfair, if it is **consistently** weighing **relevant** financial criteria. If the algorithm recommends loans to one group of users, but denies loans to another set of nearly identical users based on unrelated/irrelevant criteria, and if this behavior can be repeated across multiple occurrences, an algorithm can be described as biased.

The important point here is that the relevant criteria for making a decision are the only ones that should count in making that decision

Algorithmic Bias

To illustrate this point consider the following:

A ruling by the European Court of Justice in 2011 requires that to eliminate gender discrimination in setting insurance rates, insurers must not give lower premiums to women drivers, (or better pensions to men because they have shorter life spans).

But stats show that women are in fact less likely to have accidents, and insurances companies work on assessing risk.

So any decision making process that works out premiums based only on relevant factors and not on protected characteristics (gender, ethnicity, sexual orientation etc) may end up giving more favorable premiums to women.

But arguably that would **not** be inappropriate ?

Algorithmic Bias

It would be inappropriate (unfair) if the ML algorithm *uses the gender of any individual person as a factor in deciding the premium*, since a man may satisfy all the *relevant* criteria for a lower premium, but be given a higher **premium because he is a man**

The point is that one can acknowledge that statistically women have less accidents then men, but the gender of a person is not in and of itself relevant to deciding whether any given individual is a safe driver. Other factors are relevant, such as the driving history of that individual, the driving experience of that individual, the kind of car that individual is requesting insurance for etc etc.

Now, it maybe that an outcome of using an ML algorithm is that women are given lower insurance premiums, but if the ML used only the appropriate relevant factors in deciding premiums for any given individual, then arguably the ML is not *inappropriately* biased (even though the European Court of Justice state that the outcomes of such algorithms should not favour women)

Types of Algorithmic Bias

Data Bias if the original training data is biased, the algorithm will perpetuate and potentially compound these biases.

A key problem with machine learning algorithms trained on data sets that reflect societal biases at the time of data collection. But society evolves and changes so that what might not be considered biased/unfair at a given moment in time (at the time when the training data was collected) might later be considered biased/unfair. How then do we make sure that the ML training data sets are appropriately updated in line with the way society evolves ?

Data Bias relative to statistical standards

In general, training data may not be representative enough for anticipated, or even unanticipated contexts of use. The biases in these cases are relative to a statistical standard (e.g., the statistical features of more diverse geographic areas).

E.g. If training data for autonomous vehicle (AV) based on one particular area (e.g. London), but then AV deployed in areas that don't share same characteristics (e.g. rural areas).

Types of Algorithmic Bias

Data Bias relative to moral standards

Sometimes, biases are relative to moral standards.

E.g. suppose an algorithm allocates money for health spending, based on which treatments are more successful, but the less successful treatment is for health problems in more under-privileged sub-populations*, and is less successful because they are under-privileged*, then (ceteris paribus – i.e., all else being equal) one **ought** not to prioritize spending on the more successful treatment

Types of Algorithmic Bias

Algorithm Bias Bias in the design of an algorithm. Eg, a search engine that shows three results per screen can be understood to privilege the top three results slightly more than the next three

Use and Interpretation Bias The use of an algorithm's output can also be subject to bias resulting from:

i) Use in unintended contexts, e.g. AV knowingly used in environments for which it was not trained
ii) Interpreting outputs as objectively correct, e.g. a probation officer who uses algorithm to predict risk of re-offending, and believes that the algorithm is entirely objective and infallible, and automatically accepts suggestions without considering alternative assessments which she could draw from her own knowledge and experience.

Feedback Bias Many algorithms further trained and optimised over time, usually based on data concerning whether the original recommendations or predictions were judged accurate or useful by human operators. This means that any of the above biases can be further amplified over time

Recent Examples of Bias

The explosion of machine learning based AI has made algorithmic and data bias a particularly important issue

Here are some recent examples:

1. Lock them up and throw away the key

COMPAS algorithm widely used in US to guide sentencing by predicting the likelihood of a criminal reoffending. In 2016 it was revealed that COMPAS is racially biased. An independent news organisation compared actual re-offending rates against the system's predictions and showed that COMPAS predicts that black defendants have a higher rate of reoffending then they actually do, and the reverse for white defendants.

(https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm)

Recent Examples of Bias

2. The criminal minority

In several US states, PredPol algorithm predicts when and where crimes will take place, with the aim of helping to reduce human bias in policing. But in 2016, the Human Rights Data Analysis Group found that PredPol unfairly targets certain neighbourhoods

When researchers applied a simulation of PredPol's algorithm to drug offences in Oakland, California, it repeatedly sent officers to neighbourhoods with a high proportion of people from racial minorities, regardless of the true crime rate in those areas, because the software learns from reports recorded by the police rather than actual crime rates, so creating a *feedback* loop that can exacerbate racial biases

https://www.newscientist.com/article/mg23631464-300-biased-policing-is-made-worse-by-errors-in-pre-crime-algorithms/

Recent Examples of Bias

3. Here's looking at you white man

Facial recognition software is increasingly being used in law enforcement and is another potential source of race and gender bias.

In 2018, researchers at the MIT found that three of the latest gender-recognition AIs, from IBM Microsoft and Chinese company Megvii, could correctly identify a person's gender from a photograph 99 per cent of the time – but only for white men. For dark-skinned women, accuracy dropped to 35% !

This increases the risk of false identification of women and minorities. Again, it's probably down to the data on which the algorithms are trained: if it contains way more white men than black women, it will be better at identifying white men.

https://www.newscientist.com/article/2161028-face-recognition-software-is-perfect-if-youre-a-white-man/

Recent Examples of Bias

4. Facebook feeds the Intifada

Sometimes AI feeds back to heighten human bias.

In October 2017, police in Israel arrested a Palestinian worker who had posted a picture of himself on Facebook posing by a bulldozer with the caption “attack them” in Hebrew. Only he hadn’t: the Arabic for “good morning” and “attack them” are very similar, and Facebook’s automatic translation software chose the wrong one. The man was questioned for several hours before someone spotted the mistake. Facebook was quick to apologise.

Some Problems with Identifying Algorithmic Bias

Complexity The complexity of analyzing algorithmic bias has grown alongside the complexity of programs and their design.

Decisions made by one designer, or team of designers, may be obscured among the many pieces of code created for a single program; over time these decisions and their collective impact on the program's output may be forgotten.

The Black Box problem Explaining how Machine Learning algorithms make decisions is a major unsolved issue – *the black box problem* (their *opaqueness*) - even for AI researchers and developers.

Solutions to the Problem of Algorithmic Bias

Joint Google and Microsoft working group named *Fairness, Accountability, and Transparency in Machine Learning.*

Ideas from Google include community groups that patrol the outcomes of algorithms and vote to control or restrict outputs they deem to have negative consequences.

In recent years, the study of the Fairness, Accountability, and Transparency (FAT) of algorithms has emerged as its own interdisciplinary research area with an annual conference called FAT (Fairness, Accountability and Transparency) https://fatconference.org/

Solutions to the Problem of Algorithmic Bias

Summary of key techniques for solving problems with bias

1. **Statistical Approaches:**

- *Pre-processing* involves modifying training data, with the aim of preventing algorithm from learning discriminatory decision-making rules in training stage.
- *In-processing* involves modifying algorithmic model itself.

E.g. change the criteria that result in 'branches' in a decision tree in order to ignore or correct influence of **protected characteristics** (gender, ethnicity, sexual orientation etc)

However, many in-processing methods require personal data regarding protected characteristics, which may not be available due to the legal sensitivity of data

Solutions to the Problem of Algorithmic Bias

- *Post-processing* involves removing discriminatory rules or otherwise modifying a model (e.g. confidence intervals, weights, probabilities, predicted classes or labels) after it has been trained.

Eg, modifying a model so it places less significance on particular postcodes, which could be closely correlated with one specific ethnic group. Outcomes or decisions can also be artificially adjusted to ensure equal treatment across groups within the affected population. For example, if it is known that a probation risk assessment algorithm consistently ranks one ethnic group as a higher risk than others, any risk assessment relating to an individual from that group might be downgraded by a human probation officer to ensure an equitable outcome.

Software toolkits are now being developed which encompass these statistical methods for measuring and mitigating bias in algorithmic decision-making systems. Two specific examples are

Accenture's 'Fairness Tool', and IBM's 'AI Fairness 360 Open Source Toolkit'

Solutions to the Problem of Algorithmic Bias

2. Discursive frameworks, self-assessment tools and learning materials:

A number of tools have been developed for self-assessment, education, and interaction with stakeholders affected by algorithmic systems.

Eg, the US-based *AI Now Institute* has proposed 'Algorithmic Impact Assessments (AIAs)', a self-assessment framework for public agencies to assess the potential impact of automated systems (similar to existing environmental and data protection)

Solutions to the Problem of Algorithmic Bias

Key Elements of AI Now's Algorithmic Impact Assessment

1. Agencies should conduct a self-assessment of existing and proposed automated decision systems, evaluating potential impacts on fairness, justice, bias or other concerns across affected communities
2. Agencies should develop meaningful external researcher review processes to discover, measure or track impacts over time
3. Agencies should inform public, of their definition of "automated decision system," and any related self-assessments and researcher review processes before the system is deployed
4. Agencies should invite public comments to clarify concerns and answer outstanding questions
5. Governments should provide mechanisms for affected individuals or communities to challenge inadequate assessments or unfair, biased or otherwise harmful system uses that agencies have failed to address.

Solutions to the Problem of Algorithmic Bias

3. Documentation standards:

Big data analytics and machine learning facilitated by the usage, sharing and aggregation of diverse datasets, sometimes with only a limited understanding of how this data has been generated and what its strengths and weaknesses are.

Hence risk that unintended biases will be introduced to the datasets or the way in which they are processed by, for example, unintentionally using a dataset without understanding its original context. Hence, efforts to standardise documentation which comes with datasets.

Documentation standards are intended to establish the basic information to be filled out when collecting new data or training a new model, which can then inform the decision-making of other developers and researchers in the future. Such information could include the creation, contents, intended uses, and any relevant ethical and legal concerns about the data.

This information would help users interrogate datasets and identify potential biases in datasets and models prior to and during processing.

Solutions to the Problem of Algorithmic Bias

4. Technical Standards and Certification Programmes:

A number of national and international organisations have started to publish technical standards which could help mitigate algorithmic bias in the design and deployment of AI systems.

E.g., the Institute for Electrical and Electronics Engineers (IEEE and its standards association IEEE-SA)

IEEE P7003TM Standard for Algorithmic Bias Considerations (http://fairware.cs.umass.edu/papers/Koene.pdf)