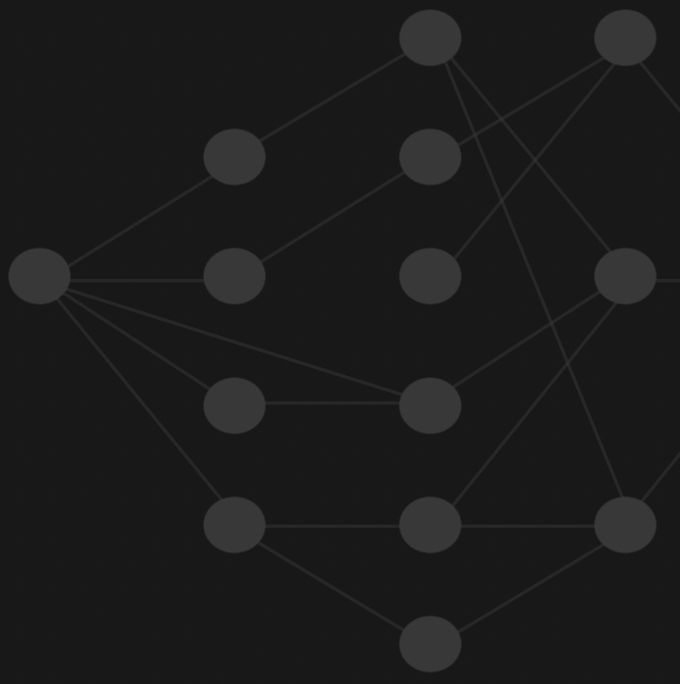
Hyperscalar Research

Bias, Fairness and Ethics in AI

Reviewing the state of the art

By:
Hema Krishnamurthy
Research Consultant, Hyperscalar
&
Tony Kenyon
Chief Scientist, Hyperscalar



V1.0 September 2020

Contents

01 | Introduction

Data-driven innovation is unquestionably a huge success, however as these advanced systems are rolled out, and they become increasingly important in decision making, we are starting to see some worrying side effects. Today we see increased concerns around *bias*, *fairness* and *ethics*, and in some contexts (especially those where decisions affect people) these are fast becoming contentious topics.

FAT (Fairness, Awareness and Transparency) has gained a high momentum in the recent times, stemming from various real-world instances of data collection without consent, enhanced monitoring of citizens, misuse of data etc.

To offer a low friction human interaction, most AI training models are built as a black box, but this has in turn led to the problem of how to ensure algorithmic transparency, explainability, auditability and accountability i.e. how decisions are reached is core to understanding the black box problem. We note that AI algorithms are subject to the *transparency paradox* i.e. they aren't always opaque intentionally.

There is great deal of literature around various ethics models [19] and hence in this paper, we focus more on what introduces **bias** in an ML system and how **fairness** can be achieved. We will touch on ethics and provide appropriate references throughout.

Some definitions

Before we get too deep into this paper let's first define the main subjects, as there is often confusion around the scope of each:

Bias: From a statistical viewpoint, bias is defined as the deviation from a state of truth. Machine learning algorithms that discriminate against particular individuals, subgroups or groups for example are declared to have bias. Bias can be introduced in various ways and is not simply limited to the algorithms used, it may originate all the way back into the data, or even the data collection method, as we discuss later. Cathy O'Neil, a data scientist, calls biased algorithms 'Weapons of Math Destruction.' The problem, as she puts it, is that algorithmic models are generally opaque or incomprehensible to most users and where these models incorporate bias, this can be replicated and compounded on a much large scale and can be difficult to analyse.

Fairness: The notion of fairness in AI is quite complicated (often used interchangeably with *bias*), more difficult to articulate definitively, and the semantics are hard to pin down. However, there are different criteria to determine fairness, which we discuss in later sections of this paper. Fairness can be achieved at an individual level or at a group level (note that group fairness doesn't imply individual fairness). Recent work on collating various definitions on fairness, as well as statistical measures, and differences between actual and expected outcomes, and causal reasoning can be found in [3]. Deciding what is 'fair' may depend

heavily on the context, and what is determined to be fair now, may not be true in the future.

Ethics: The word *ethics* is derived from the Greek word *ethos*, which means habit or custom. Ethics is a branch of philosophy that aims to distinguish between right and wrong. However, this is sometimes a grey area since in some cases what one considers morally right may not be so for another. Ethics has traditionally been described from different perspectives – deontological (universal ethical rules), teleological (focus on outcomes and consequences of the actions), virtue ethics (trolley problem – a person's moral reasoning). The trolley problem [20] was a thought experiment by Philippa Foot, a virtue ethicist to show the problem with the first two models. The trolley dilemma has consequences in various AI applications like autonomous cars, automation of jobs etc.

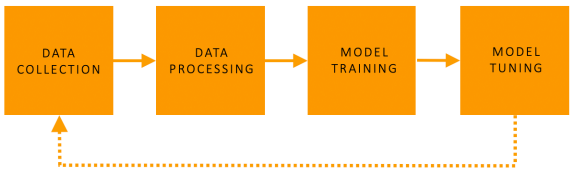
Explainability: Explainability in AI (XAI) is defined as the methods used to understand and unequivocally interpret the decisions or predictions made by ML models. This is an emerging field focussed on opening up the classic 'black box' view of ML. It aims to provide a means to interpret, understand and trace why ML models made the decisions they did, thereby offering more trust in those decisions, and the ability to diagnose where decisions and predictions have gone astray.

Drift: *Concept drift* refers to the change in the target variable (that is being predicted) over time, whereas *data drift* refers to the independent variables (input data) themselves changing over time. We need to pay close attention to concept drift because this essentially means that predictions or decisions being made by the model may start to become adversely decoupled from the true state of the environment. We discuss this topic in more detail later.

There are several questions around bias and fair outcomes. Does privacy equate to fairness? [4] When is an algorithm said to have a bias?

02 | Types of bias in AI

A typical machine learning workflow comprises several steps, and bias can creep in at any stage. Simplistically these steps are illustrated in the following diagram, however, note that this flow is typically highly iterative, and models are rarely static:



The dynamic nature of AI workflow and data means that even good models can 'drift' to behave badly over time, and this is something we discuss later in this paper.

Bias introduced in data collection

Human bias is induced based on what type of data is collected or the data collected not being representative of the target population. Protected attribute data needs to be collected to assess fairness. For the purposes of our discussion, note that protected attributes are attributes or characteristics that lend themselves to be a source of intentional discrimination.

Bias could also be introduced by using historical data that perpetuates stereotypes. For example, if we design a recruitment system trained on data from existing employees, at first glance this may seem a reasonable approach, however such a model may be subtly weighted against candidates from a more diverse background.

Bias introduced in data processing

In pre-processing there is often a temptation to over clean the data, and this should always be handled with extreme care, with the full knowledge of what the data represents and the application domain. For example, outlier removal can be extremely helpful in improving model performance but could significantly skew analysis in fields such as clinical trials (making any results unusable).

In supervised learning, data typically needs to be carefully labelled, and this can be achieved in various ways (some automated, some manual, possibly even outsourced (e.g. Amazon Mechanical Turk [29]). Labelling can be a subjective process and if so, may heavily influence learning algorithms during training, and this can lead to significant bias in production models.

Where data is hard to source this can be a particular problem, as we me have to create models that are trained on labels that represent only part of the production use cases. This can amplify exiting bias or bias the whole model in production.

Algorithmic bias

As we have seen, if earlier decisions in data collection or processing are subject to some form of prejudice, this bias can ripple up to influence algorithmic decision making, compromising the

whole system. Algorithms themselves can reinforce and amplify bias inherent in the data, or perhaps even introduce new ones.

It is also important to note that algorithmic bias isn't always intentional. Bias can be introduced through unconscious choices made by algorithm designer, at various stages. For example:

1. Model selection and development: Bias can be compounded if an incorrect algorithm that doesn't suit the application is chosen.

2. Model start conditions: The starting conditions for a model often can be changed to achieve better outcomes.

3. Model tuning and evaluation: Tuning the parameters changes the model; selection of performance metrics and thresholds can affect bias.

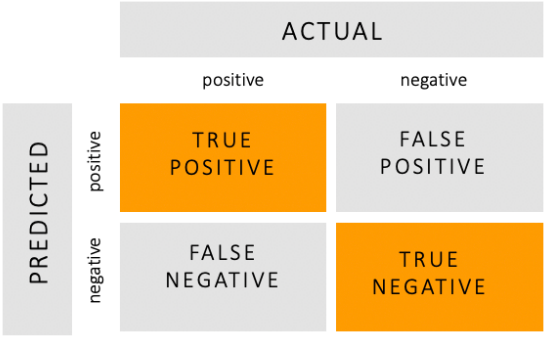
Machine learning tends to automate bias, but is necessarily always a bad thing? The authors of [7] argue that this may not always be so, and that certain types of bias – such as domain expertise - may lend themselves to creating well-suited models for a given application.

03 | Types of fairness in AI

A number of criteria for algorithmic fairness have been proposed [3], and we discuss these below:

Fairness through blindness entails removal of protected attributes. But protected attributes whose labels are deliberately removed may correlate with other features in the remaining training data. Researchers call this "redundant encodings" – when membership in a protected class is also encoded in other data. Examples are the Google ad-listing tool and Amazon resume tool, which demonstrate that this approach promotes bias if the model relies on data that contains historical bias. Hence this approach is fundamentally flawed.

Demographic parity: This refers to the approach of creating groups and requiring statistical parity on the outcomes on these groups. It is meant to promote group fairness but sometimes fails to address individual fairness. The predicted outcome is equalized across protected attributes of the entire set. There is no pre-set outcome defined for this.



Equalized opportunity: The *true positive* rates are set to be the same for the protected group versus others. Here, the equalisation is performed only on the subgroups, which have a positive value of the outcome.

Equalized odds: Training is done on data for which outcome Y is known with certainty. This is similar to demographic parity, but instead equalises across subsets with the same outcome Y i.e. as opposed to equalized opportunity, this approach extends fairness to both those with positive and negative outcomes. In this case, odds are equalised amongst *true positive* and *false negative* for diff values of the protected attribute. Thus, equalized opportunity and odds are more useful if accuracy is key.

Counterfactual fairness: Data is altered to place an individual in one group when in reality they belong to the other. The trick here is when the protected attribute is flipped to the counterfactual value, the change must be reflected in other variables dependant on the protected attribute. However, this approach is hardly prescriptive.

04 | Bias & fairness in ML algorithms

In this section we discuss the bias-fairness issues that can occur with widely used ML techniques and algorithms:

Supervised learning

K-Nearest Neighbours (KNN): A data element is assigned to a class, which has the largest number of k nearest neighbours. However, this could lead to new elements being put in the better-represented bucket, thus introducing unfairness.

Linear Regression: Prediction is made by computing a weighted sum of the input features plus a constant bias term, and the fairness assessment depends heavily on the assigned weights and the bias term. Unfairness may be introduced into the algorithm if the weights aren't picked objectively.

Logistic regression: A feature vector is mapped to a class label, and therefore bias can creep in during the class-labelling phase (which can be a subjective exercise).

Support Vector Machines (SVM): SVMs use hyperplanes (decision boundaries) to classify input data. The goal is to find classifiers with as big margins as possible. SVMs are prone to implicit bias from the training data towards protected attributes.

Semi-supervised and unsupervised learning

Semi-supervised learning uses data sets that are partially labelled (E.g. Google Photos) and is a combination of supervised and unsupervised techniques.

Principal Component Analysis (PCA) is used for dimensionality reduction, which comes with a cost of reconstruction loss. When reconstructing data, information could be lost leading to unintended bias creeping in.

Adversarial learning wherein two learners, one predicting the output, the other the protected attribute - in order to converge on a model that predicts the correct outcome independent of the protected attribute, has been shown to reduce bias [8]. There has been work in the area of adversarial networks to treat sensitive attributes in the data as *nuisance parameters*. In statistics, nuisance parameters are ones that are not of immediate interest but must be taken into account for statistical analysis [9].

Note that fairness is easier to define for supervised learning than for unsupervised settings because supervised learning is performed with intent to predict an outcome, versus unsupervised learning which is typically used to gather insights on the data, perform anomaly detection, or perform associative rule learning.

05 | Explainable AI (XAI)

Explainability, closely tied to traceability and transparency, is important in understanding bias, fairness and the ethical use of an AI mode. Explainability is challenging, often subjective and done *post facto*, so ideally, we want to take the human out of the loop. According to Google researcher Peter Norvig:

"*You can ask a human, but, you know, what cognitive psychologists have discovered is that when you ask a human you're not really getting at the decision process. They make a decision first, and then you ask, and then they generate an explanation and that may not be the true explanation*."

In many cases, it can be problematic to unpack AI decision making, particularly with black box models such as neural nets. High explainability often comes at the cost of accuracy and performance. Simpler ML techniques, such as decision trees, Bayesian classifiers etc., provide greater visibility into decision paths, however with more complex models such as deep neural networks, the problem of explainability can be extremely challenging.

Unless techniques improve quickly it is possible that the choice of ML model may be influenced by the use case, for example where there is a need to comply with laws and regulations - such as GDPR Article 13, "Right to Explanation" [28].

XAI is being explored from several perspectives (for example some commercial tools provide a form of explainability by exhaustively testing all input variants). There are open source tools to assist, such as *AI Explainability 360* by IBM [10] that are being evaluated. There remain open questions around whether explainable discrimination is still acceptable [12].

We believe for the time being more effort should perhaps be focussed on analysing bias and fairness. This is an area of active research, beyond the scope of this paper. For further information Google have produced a very useful white paper [27].

06 | When AI goes wrong

Here we highlight some controversial and well-publicised examples where bias, unfairness and ethical issues introduced in AI models has become highly problematic. Part of the challenge here is the potential difficulty in analysing and exhaustively testing model behaviour prior to deployment. Although tools in this area are evolving rapidly (and we provide some useful links shortly), this remains an active area of research.

COMPAS – This is one of a number of risk assessment tools used in the US criminal justice system. The COMPAS algorithm is designed to assist judges in deciding whether a defendant should be kept in jail or released while awaiting trial, by providing a *risk score*. This tool is designed to remove a judge's intuition and bias. However, a ProPublica report in 2016 found that this algorithm was still biased, finding that in some contexts black defendants were more than twice as likely to be labelled as high risk than white defendants [13]. Further analysis published in the MIT Technology Review [14] illustrates the problem clearly and demonstrates ways to mitigate the underlying bias.

PREDPOL - Predictive policing uses algorithms to analyse police data, sometimes combined with other types of government and commercial data, to identify patterns and make predictions about where crime might occur, or who might commit a crime. However, this has led to reinforced racial bias [4,5]. PREDPOL's proprietary algorithm primarily uses crime type, location and timestamp information to create a map of predictive spatial hotspots but studies point out that the algorithm tends to introduce negative feedback loops [22].

When a feedback loop occurs in this case i.e. the model's predicted outputs are reused to re-train the model repeatedly, it leads to police officers getting repeatedly sent to certain neighbourhoods – typically ones with a high number of racial minorities, irrespective of the true crime rate in the area. This leads to inherent bias in the data getting reinforced.

Problems of algorithmic bias and discrimination are not new:

St. George's Hospital, UK – back in 1988 developed a set of algorithms to help filter medical school applicants based on prior admission decisions [18]. Unfortunately, this labelled dataset (the 'ground truth') perpetuated racial bias that had been systemically built into the data from previous years [17]. The British Medical Journal noted at the time, *"[T]he program was not introducing new bias but merely reflecting that already in the system*".

As an example of ML ethical violations, perhaps the most notorious in recent times is as follows:

Cambridge Analytica – In 2018 Cambridge Analytica harvested personal data from millions of Facebook users via an app - without consent. Cambridge Analytica sought to sell the data of American voters to political campaigns, for use in political advertising. Whilst these actions were totally unethical, this was also classed as a major data breach - the largest known breach in

Facebook history. The data breach was disclosed in 2018 by Christopher Wylie, a former Cambridge Analytica employee. Facebook subsequently apologised for their role in the data harvesting and CEO Mark Zuckerberg testified in front of Congress [22].

Deepfakes – Deepfakes are synthetic images, video, and audio, most often attempting to fake people and their actions; typically created using Generative Adversarial Networks (GANs). In some contexts, they can be extremely damaging to individuals, and as such should be viewed as unethical. Deepfakes are especially problematic as the results become more realistic and difficult to differentiate from reality, and they are often distributed virally on social media platforms. Several big companies (including Facebook) are now taking positive steps towards banning deepfakes. According to their blogpost [23], for an image to be taken down, one of the criteria is "*It is the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic.*" While generative AI has its place in various applications, and not all applications of deepfakes are harmful, there is a clear tendency at present for misuse.

Google Duplex - In a move to real time supervised learning, Google Duplex was built to place calls on behalf of humans and successfully complete conversations and tasks on its own (placing appointments, making reservations etc) without any intervention from a person on Google's end for the most part. However, several ethical concerns arise around impersonation, identity theft when AI speaks on behalf of humans [24].

Finally, we would be remiss if we didn't broach the subject of ML for forecasting and patient care in the context of the COVID-19 pandemic. Even here there have been ethical concerns around the *intended* use of patient data (for example, with early trials of the UK contact app), and it remains to be seen how fair these algorithms are [11].

Even good models can go rogue

Whilst we have talked a lot about bias and fairness being introduced to new models, it's also important to understand that even good models can be subject to concept drift over time. This typically happens where models are continuously fed new data and retrained or retuned. New training data may start to introduce bias if there are changes in the way it is collected or processed. It is possible that good decision-making can be influenced and degraded over time, based on these kinds of changes, so periodical testing and analysis should be performed to assess whether there is significant drift.

One way to deal with concept drift is to re-label old data and re-train the model. As you might imagine, this is not always straightforward, for example if live production data has also changed in character (legitimately) over time in response to seasonal trends in purchasing behaviour. We also need to be aware of the concept of data drift, where the data itself may have moved from expected or acceptable bounds, and to fix data drift, new data covering new classes may need to be added to the training set and the model re-trained. Drift can be very challenging to diagnose and address and often requires intimate knowledge of the data, the domain and strong analysis of the model outputs.

Tools to evaluate bias & fairness

Some tools that are available to evaluate fairness are highlighted below:

- Microsoft FairLearn tools, see [25]
- Google What If Tools, see [26]

07 | Regulations

As AI becomes ever more entrenched in decision making, we should expect regulation to play an important role in providing safeguards and methods of recourse. Whilst lawmakers are finding it hard to keep up with these advances, we are starting to see complementary regulations.

For example, in 2019 the US Congress drafted legislation to regulate AI (the *Algorithmic Accountability Act* which requires big companies to audit ML systems for bias and discrimination), with several countries drafting similar legislation shortly after [15].

We should also keep in mind that data privacy has a large part to play in machine learning and data science, with regulations such as the European GDPR act, which places strict responsibilities on holders and processor of PII data, and the need for 'informed consent' and a "right to explanation" [13].

There have been several public examples of data being used in ML models without user consent (such as Cambridge Analytica, discussed earlier, in in the healthcare sector), which raise important ethical and privacy concerns.

08 | Conclusions

AI is evolving rapidly and there is huge enthusiasm to implement and supplement systems that can leverage the significant benefits. However, we need to tread carefully, given the relative immaturity of this space, and the potential from harm in decision making - particularly where AI impacts people. According to Elizier, 2008: "*By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it*" [16]. As stated in [16], this becomes even more important "When something is universal enough in our everyday lives, we take it for granted to the point of forgetting it exists." Therein lies the danger.

Arguably one of the toughest challenges data scientists are facing today is how to demonstrate and guarantee fairness, when machine learning algorithms are evolving, models are being continuously re-tuned, and new data is being fed in. It is especially challenging in many cases to explain why decisions were made for all possible inputs (especially where there is high dimensionality on those inputs). If we are going to treat data as an asset class, particularly in highly regulated contexts, then the algorithmic systems that work on this data would need to be subject to strict controls and transparency. How should data be collected, used, governed, managed? Will the ethical frameworks need to be reworked?

From a philosophical perspective, ethics can be an abstract concept, but when it comes to designing ML algorithms in the real world, serious concerns emerge on how practitioners should design algorithms and models that clearly differentiate between right and wrong, and in understanding how bias can be introduced systemically in various, often subtle ways. Hence data and algorithm ethics must be guided by ethical frameworks to ensure responsible innovation, and that these models be built to maximise accuracy – but with appropriate constraints.

With algorithms emerging as a powerful tool of social control, whilst algorithmic fairness may not be 100% achievable for every application, this should not be a reason to obstruct the significant benefits of AI – what we do need are appropriate tools, regulations and frameworks to ensure we stay on the right path.

References

1. Songül Tolan, Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges (2019)
2. Sahil Verma, Julia Rubin, See: Fairness Definitions Explained (2018)
3. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Rich Zemel, Fairness through Awareness (2011)
4. https://www.predpol.com/law-enforcement/
5. http://www.abnms.org/uai2012-apps-workshop/papers/ChoEtal.pdf
6. Tom M. Mitchell, The Need for Biases in Learning Generalizations
7. Christina Wadsworth, Francesca Vera, Chris Piech, Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction (2018)
8. Gilles Louppe, Michael Kagan, Kyle Cranmer, Learning to Pivot with Adversarial Networks (2017)
9. https://arxiv.org/pdf/1611.01046.pdf
10. https://aix360.mybluemix.net/
11. https://healthitanalytics.com/news/machine-learning-tool-predicts-staffing-needs-during-covid-19
12. https://link.springer.com/chapter/10.1007/978-3-642-30487-3_8
13. ProPublica report on COMPAS bias, 2016. See: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
14. “Can you make AI fairer than a Judge?”, MIT Technology Review, June 2020.
15. “Congress wants to protect you from biased algorithms”, MIT Technology Review, April 2019.
16. Yudkowsky, E., "Artificial Intelligence as a Positive and Negative Factor in Global Risk", Machine Intelligence Research Institute. 2008
17. Barocas, S. and Selbst, A.D., 2016. Big data's disparate impact. Calif. L. Rev., 104, p.671.
18. Lowry, S. and Macpherson, G., 1988. A blot on the profession. British medical journal (Clinical research ed.), 296(6623), p.657.
19. Ilona Baumane-Vitolina, Igo Cals, Erika Sumilo, 2016. Is Ethics Rational? Teleological, Deontological and Virtue Ethics Theories Reconciled in the Context of Traditional Economic Decision Making.
20. Judith Jarvis Thomson, 1985. The Trolley Problem.
21. Cambridge Analytica on Wikipedia: https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal
22. Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, Suresh Venkatasubramanian, 2017. Runaway Feedback Loops in Predictive Policing
23. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/
24. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html
25. Microsoft’s FairLearn tools: See: https://docs.microsoft.com/en-us/azure/machine-learning/concept-fairness-ml
26. Google What If Tools: https://pair-code.github.io/what-if-tool/ai-fairness.html
27. Google White Paper “AI Explanations”: See: https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf
28. GDPR Article 13, "Right to Explanation": https://gdpr-info.eu/art-13-gdpr/
29. Amazon Mechanical Turk. See: https://www.mturk.com/

About Hyperscalar

Hyperscalar provides research and advisory and IP diligence functions, specialising in private equity, as well as advising and mentoring start-ups in disruptive technologies such as machine learning, blockchain, process automation, cybersecurity and robotics. For further information see **www.hyperscalar.com**