The Aletheia Framework™
2.0

USING THE AI BIAS ASSESSMENT TOOL

ROLLS ROYCE

What is bias in artificial intelligence?

Bias is a disproportionate weight in favour of or against something or someone. In an AI context, it's critical to understand if a biased output is objective, or the result of prejudice at some stage in design, development or operation of the algorithm.

Bias can occur anywhere in the AI lifecycle. This could be because of bias in the original brief, or because the data is biased – due to sensor or human sampling errors, for example. Or it could be because the underlying algorithms are biased – reflecting a conscious or unconscious prejudice from its developer, or because it's been trained using biased data.

> Bias is not always bad, which is why we need to understand its impact, likelihood and detectability.

Should we eliminate all bias?

Bias is not always wrong. For example, younger people will appear most in a dataset of fastest 100m running times – it‘s biased, but an accurate reflection of reality. Similarly, safety margins are a form of 'positive' bias – riding an autonomous bicycle 'as close as possible' to the edge of a cliff, you'd want to know it has a 2m safety bias built-in.

So, it's not enough to simply identify the existence of bias. The more pertinent question is the impact of that bias, how likely it is to occur, and whether you can prevent or detect it. A successful strategy for managing and eliminating bias must be based on fully assessing these factors, to ensure your AI is ethical, safe and trustworthy.

What's the best way to manage bias?

In engineering, there is a risk management tool that approaches potential failures in exactly the same way – i.e. factoring their impact, the likelihood of that failure occurring and whether you can detect it, and calculating a total risk score based on this data. The tool is Failure Mode Effects Analysis (FMEA), and it's the gold standard methodology for assessing risk in novel and complex systems, used by Rolls-Royce for decades.

The Aletheia Framework AI bias assessment tool is modelled on the FMEA approach, reengineered for assessing artificial intelligence.

What makes the FMEA approach so powerful?

For developers, it‘s a more practical approach that results in a more complete understanding of bias risk in their AI. It also makes it easier to identify potential bias. Detecting bias by examining outputs is hard and often not appropriate or useful. Detecting bias by examining its potential sources – as the tool does – is both easier and more useful. Bias can be pre-emptively managed or eliminated, resulting in better, more ethical AI, and saving the developer time.

> The FMEA approach looks beyond the algorithm to its broader context, making it easier to identify and eliminate harmful bias.

Where can AI bias occur?

Bias can occur in most stages of the AI lifecycle, from design and development through to deployment and use.

DESIGN →

Developing the brief for a new AI

Potential sources of bias:

- The stakeholder requirements that inform the brief are biased

DEVELOPMENT →

Building, training and testing the AI

Potential sources of bias:

- The initial dataset used is biased, or is cleansed in a way that introduces bias
- The model is biased, through poor selection, bias in algorithms, or bias in neural network weightings
- The data used to train the model is biased
- The tuning/optimisation of the model is biased
- The test data is biased

DEPLOYMENT

Rollout & ongoing operation of the AI

Potential sources of bias:

- The developed AI is biased and propagates into production
- The operational data acquisition method is biased
- The data transfer mechanism introduces bias
- Data operations or manipulation introduce bias

Using the AI bias assessment tool

STEP-BY-STEP

1

Thinking through each stage of the design, development and deployment of an AI, consider the 'types' of bias and whether they could be present in your case (using the list of potential sources on the previous page as a starting point).

2

Using the 1 – 10 scoring criteria at the end of this document;

a. Assess each of these for severity of impact

b. Then, consider the potential causes and assess each for the likelihood of it occurring. In considering the causes, it's important to ascertain the root cause – for example, if the cause of a failure is that training data is biased, keep asking "why?" to get to the root of that bias

c. Finally, assign a score for ease of detection or prevention, based on the current plan or deployment.

3

This will give you a total Risk Priority number (on a scale of 1 – 1,000). Starting with the highest scores, create preventions to mitigate the impact or likelihood of the bias.

4

Using the columns at the end of each row, re-score each step with the mitigations taken into account.

Use the AI bias assessment tool ⟶

Note – The AI bias assessment tool is intended to provoke thinking. It identifies a process for assessing bias risk, and potential sources that a developer should consider, but it cannot be exhaustive as every AI application is different.

EXAMPLE #1: MITIGATING AN ETHICAL IMPACT

Consider the development process of an AI to sift CVs and select candidates for interview.

If the data being used to train the AI has been synthesised by one person, that AI is very likely to be biased.

OCCURRENCE – ALMOST CERTAIN = 10

The impact would be a biased selection of candidates for interview – a complete failure in the AI's task to select the 'best' candidates.

SEVERITY – CRITICAL = 10

There is no means of detecting this bias in the data set by examining the outputs.

DETECTION – ALMOST IMPOSSIBLE = 10

This leads to a **total risk priority score of 1,000** – the highest possible.

In this case, a simple mitigation can be applied. It's easy to identify the root cause – that only one person is synthesising the data. The issue can therefore be mitigated by ensuring a diverse team synthesises the data used to train the AI.

This mitigation changes the Detection score by monitoring the cause (the team doing the synthesising) rather than the output (the candidates who have been selected).

DETECTION – ALMOST CERTAIN = 1

… and it means there is no longer a 'very high' chance of the bias occurring.

OCCURRENCE – VERY UNLIKELY = 2

Revised total risk factor score (Severity x Occurrence x Detection) = 20
It becomes an unlikely source of bias that does not require further mitigation.

EXAMPLE #2: MITIGATING A BUSINESS IMPACT

Consider the development of a new inventory management system.

If the new system is designed by the designer of the old system, it will likely be biased towards similar technologies and data sources.

OCCURRENCE – HIGHLY PROBABLE = 8

Although the impact wouldn't necessarily be unethical, it would very likely be suboptimal having a high impact on business metrics.

SEVERITY – VERY HIGH = 8

There is no means of detecting this bias by examining the design.

DETECTION – ALMOST IMPOSSIBLE = 10

This leads to a **total risk priority score of 640** – very high.

In this case, a simple mitigation can be applied. It's easy to identify the root cause – that only one person designed the new management system and they also designed the one being replaced. The issue can therefore be mitigated by ensuring a diverse team creates the new system.

This mitigation changes the Detection score by monitoring the cause (the team developing the solution) rather than the output (the solution itself).

DETECTION – ALMOST CERTAIN = 1

… and it means there is no longer a 'very high' chance of the bias occurring.

OCCURRENCE – VERY UNLIKELY = 2

Revised total risk factor score (Severity x Occurrence x Detection) = 16
It becomes an unlikely source of bias that does not require further mitigation.

Scoring criteria

SEVERITY			OCCURRENCE		DETECTION		
Impact	Description	Rank	Likelihood of occurring	Rank	Detectability	Description	Rank
Critical	Bias causes application to completely malfunction/mislead	10	Almost certain	10	Almost impossible	No known controls available. Failure will escape	10
Extremely high	Extremely high degree of dissatisfaction due to impact of bias on downstream process or customer. Extremely high impact on ethics or business metrics	9	Very highly probable	9	Very remote	Random checks made at low frequency, no prevention in place. Failure highly likely to escape	9
Very high	Very high degree of dissatisfaction due to impact of bias on downstream process or customer. Very high impact on ethics or business metrics	8	Highly probable	8	Remote	Random checks in place, prevention based on operator noticing problem. Failure likely to escape	8
High	High degree of dissatisfaction due to bias on downstream process or customer. High impact on ethics or business metrics	7	Quite probable	7	Very low	Regular checks in place (not 100%). Prevention based on operator training. Failure likely to escape	7
Very significant	Bias causes very significant dissatisfaction, e.g. very significant impact on ethics or business metrics	6	Probable	6	Low	Prevention based on operator diligence and regular training, regular checks in place, significant chance of escape	6
Significant	Bias causes significant dissatisfaction, e.g. significant impact on ethics or business metrics	5	Might occur	5	Moderate	Failure prevention considered, 100 % checks on output, medium chance of escape	5
Moderate	Bias causes problems which have a noticeable impact on business performance metrics. Medium dissatisfaction, e.g. extra effort or rework needed or moderate impact on ethics or business metrics	4	Might not occur	4	Moderately high	Failure prevention designed into system, 100% recorded audit of process. Low chance of escape	4
Low	Bias causes minor problems which take a small amount of time to overcome. Some dissatisfaction and/or low impact on ethics or business metrics	3	Unlikely	3	High	Mistake proofing used to try and prevent failure. 100% recorded audit of process. Small chance of escape	3
Very low	Nature of bias only causes slight delay or minor rework. Slight annoyance/ very low impact on ethics or business metrics	2	Very unlikely	2	Very high	Controls nearly always prevent or detect failure. Remote chance of escape	2
Minor	The customer(s) of the process will not notice the effect of the bias. No impact on ethics or business metrics	1	Almost impossible	1	Almost certain	Controls will always prevent failure from occurring or, if it does happen, will always detect the problem before it escapes	1