The Aletheia Framework™ 2.0

FAQ

Content

Curious about The Aletheia Framework

F A Q

What is Aletheia?

It's a concise, step-by-step framework? to guide organisations through the ethical consideration of using an Artificial Intelligence application, to make sure ethics are fully considered before going ahead; and then when it is working, a process to ensure an AI's decisions continue to be highly trustworthy. It is Rolls-Royce's contribution to society's conversation around the main ethics and trustworthiness challenges of AI.

Who is it aimed at?

We anticipate it to be most useful to those developing and deploying AIs; people in organisations responsible for managing risks; ethics and compliance teams. But it is published to help build public trust in AI, so we hope that members of the public may find it helpful to see what sort of questions they can suggest organisations ask themselves when developing AI products.

How does it help?

It offers organisations that are considering using artificial intelligence a simple, clear path through the complex areas of the social impact, trust and accuracy, and governance of an AI product. Depending on the use of that AI, or at what stage it is in its lifecycle, not all 32 steps may be applicable or proportionate at a single point in time. However, the spirit of using The Aletheia Framework is to show that each step has been fully considered, and a clear rationale given for why any step might not be applicable at that time.

What is the intent of The Aletheia Framework?

We published it to help organisations struggling with turning the "what" of AI ethics and trustworthiness, into the "how". The Creative Commons licence means any organisation can take The Aletheia Framework for free and use it directly; or use it to inspire and explore building their own ethics and trustworthiness framework, with a credit to its origins and inspiration. An example of this is The Institute for Ethical AI in Education, 2021.

How does The Aletheia Framework approach job and skill loss issues through implementing AI systems?

We recognise the effect deploying AI systems could have on our employees, so it was important to us to engage with our Unions when creating principles with this in mind. By working with them collaboratively, we were able to create principles that consider not only what the human interaction with the AI systems will be, but also what affect deploying AI will have on employees and how these affects can be mitigated.

F A Q

What are The Aletheia Framework's own origins?

IOur assurance experts took the "what" of documents like the Asilomar AI Principles and EU Ethics Guidelines and, using our product safety principles in use with our jet engines, created a "how" - a process that sits in our management system.

Our principal references are:

- Good Corporation - Developing an Ethical Future for AI;
- EU Ethics Guidelines
- European Parliament Committee on Industry, Research and Energy – Draft Report on a comprehensive European industrial policy on artificial intelligence and robotics
- Asilomar AI Principles

Why didn't you use other open-source frameworks?

At the time of publishing in December 2020, no other frameworks went as far as we needed in terms of practical application, creating procedures to consider the ethics and trustworthiness of AIs.

Why do you describe it as breakthrough work?

We created The Aletheia Framework to resolve an AI challenge inside Rolls-Royce. Part of our assurance process was to have the framework peer-reviewed by an external community of subject matter experts in Big Tech, pharmaceuticals, automotive, legal and academia to compare with their own versions. To our surprise they all came back and said they hadn't seen anything as concise and practical, nor as advanced on trustworthiness. Since publishing it in December 2020, our ongoing engagement with the AI ethics community confirms this position.

In what contexts does The Aletheia Framework need to be applied?

It is context independent - we know it can be applied in our industrial technology business; and also is relevant in the non-industrial contexts on which we've been collaborating. Depending on the type of AI product being assessed against the framework, some parts may not be relevant, but it is there to cover uses of AI in critical activities in organisations.

F A Q

What assumptions have been made with The Aletheia Framework

It assumes positive intent; and that a degree of safety awareness exists in whichever organisation or individual is using it. It also assumes whoever is building an AI is doing so with clear pre-determined explanations as to why it is ethical and trustworthy. If the assessment is undertaken superficially, there will be little value gained.

How has Rolls-Royce used The Aletheia Framework?

Our management system has something called a Digital Product Integrity Passport, which is a suite of assurance checks and processes of which The Aletheia Framework is a part. We are in the process of applying this throughout our business for both internal-facing AI products relate to supply chain management, engineering product design; as well as external products and services like our Intelligent Borescope. You can read more about our case studies here.

The Aletheia Framework is able to proceduralise AI ethics in one page by providing succinct actions to turn the “what” into the “how”. Its simplicity means new AI developments can be easily incorporated into the framework in the release of updated versions – this may be more difficult in lengthier guidance lead documents.

F A Q

How can I use Aletheia?

It is designed to be used throughout the AI product lifecycle, from conception to operation. There are many different opinions on what constitutes an AI product lifecycle, you can see our version here [LINK]and how it connects to The Aletheia Framework.

What if my AI project contains intellectual property, how would I share the framework to show compliance?

The beauty of The Aletheia Framework is that it is effective without scrutinising an algorithm. It looks at the inputs and outputs of an algorithm, rather than scrutinising inside the “black box” which is always changing and often is the place where intellectual property is created. As with any IP matter, organisations must make their own judgement on what to include and what to omit, but the evidence section in The Aletheia Framework should described why there is an omission.

What type of evidence should be used to meet the principles?

Anything from prose explanations, links to existing documentary evidence, or even code libraries and links to data stores. It must be sufficiently robust evidence for the individual reviewing and signing off an AI product to assess that the ethic will be realised when the AI is deployed.

Do I need to provide evidence for all the principles?

No, the principles are there to guide the development of AI projects, not all of the principles will be relevant to every single project.

How can I ensure there is no intended bias in my system?

The Aletheia Framework v2.0 contains a new module (as part of the assurance ecosystem) that provides a process for assessing, identifying and mitigating unintended bias risk in AI requirements, algorithms, and data sets that are used in the development and use of AIs. You can see that module here [LINK].

F A Q

I don't have an IT expert in my company, how can I meet these principles?

Do what you can and make sure you explain why something may not suit your particular AI product, or be achievable at that time. For example, in the Musiio.com case study [LINK], it shows that as a start-up some steps in the process were either unaffordable or not applicable at an early stage in the company's lifecycle but could be addressed in the future.

What are process checks?

Take the example of when you might choose an energy supplier. You look at your cost of your energy supplier over the last year and see if your tariff would be cheaper with a different supplied. You're doing an independent check of the algorithm.

AI and ML tends to jump from data to interpretation, sometimes this can be an over interpretation. Often an AI isn't biased, the data is. How much does the framework cover this and how much more is there to go into this space?

Aletheia uses five process checks, these identify if the expected output of an AI has suffered from "drift" and the application no longer meets expectations or conformant (trustworthy) performance. The framework also asks developers to provide assurance for the comprehensiveness and reliability of data being used to train the AI. If the application starts to fail the process, Aletheia won't tell you what exactly has gone wrong, just that something has gone wrong. Then, as with any assurance process you have find out why it's gone wrong.

Because the reliability of training data for AIs is such a critical area for root causes of AI drift, in Aletheia v2.0 Rolls-Royce is kicking off quality assurance that looks at how to validate training data, test data and the continual feed of data. We know that just as there was a big gap on procedural AI Ethics and Trustworthiness in 2020, right now there is another big gap globally on how to do that.