
# The Aletheia Framework™

## 2.0

ROLLS ROYCE

# Foreword



It took more than just a leap of faith for the Wright Brothers to sustain the first powered flight in 1903. It was having absolute trust in the thing that propelled them skyward. Indeed, it has been our ability to not just create new technologies - but to control and master them - that has pushed humankind forward.

So is the need with Artificial Intelligence. While the use of AI technologies has already led to countless breakthroughs in the worlds of healthcare and business, it remains a tool with vast untapped potential – limited by the lack of trust many have in it.

Will the AI be accurate, true and fair? Is it being operated well, such that it leads to good outcomes for the people it is supposed to serve? Will it stand up to the same ethical scrutiny that we are expected to apply to ourselves?

As an industrial technology company that has been using AI for many years, Rolls-Royce has been working through these important questions. What's more, we've been doing that in an environment where there is a real need to trust algorithmic outcomes completely. This experience has culminated in The Aletheia Framework™ – a toolkit that we believe creates a new global standard for the practical application of ethical AI. Follow the checks and balances within it, and organisations can be sure that their AI project is fair, trustworthy and ethical. We are applying it in our business to accelerate our progress to industry 5.0.

As we emerge from the global Coronavirus pandemic, the need to free up human creativity to find new routes forward has never been greater. Making full and ethical use of our AI tools will contribute to the growth, wealth and health of our world. For these reasons, Rolls-Royce is making The Aletheia Framework™ freely available to all that might benefit from it.

**Warren East**

Chief Executive – Rolls-Royce

# The Aletheia Framework™ 2.0

The Aletheia Framework™ is a toolkit to guide the practical application of ethical AI projects. It is designed to go beyond theory to be a clear, 32-step process that any organisation can follow so that its AI is accurate, well-managed and has a positive impact on the world.

Importantly, it does not seek to influence the AI algorithms themselves, as these can be constantly learning and evolving. Instead, it provides a checklist of measures to ensure the initial design of the AI application is ethical, and that its resulting outputs remain unbiased and true to the intended design.

The Aletheia Framework™ consists of three principal areas of focus: Social Impact, Accuracy/Trust, and Governance. Each of these areas has an associated set of contexts and ethics, which ,in turn, connect directly to a series of realisation principles. It is these 32 realisation principles, or challenges that must be satisfied for the organisation to deem its AI project trustworthy and ethical.

## Social Impact

Consideration must be given to the possible impact of the AI on all potentially affected stakeholders – both inside and outside the organisation. The benefits of the project must be clearly identifiable and contribute to broader social and sustainability objectives.

## Accuracy/Trust

The AI system must be true and fair. By design, it should be safe, trustworthy and free from bias or prejudice, with sufficient checks built into processes to ensure it remains uncorrupted.


## Governance

The architecture and handling of data within the AI system must be adequately governed through planned protocols and checks. Overall security and accountability of the AI must be considered and formalised.

# Social Impact



| CONTEXT | ETHIC | REALISATION PRINCIPLES | |
|---|---|---|---|
| Benefits | AI and robotics shall be seen as delivering good. Doing good is one of the five key ethical principles of the EU guidelines for ethical AI. Good includes commercial prosperity. | 1 | Deployment of AI and robotics shall be shown to improve the well-being of employees and/or the general public, such as improved safety, working conditions, job satisfaction. |
| | | 2 | Additional to 1. (or instead of), deployment of AI and robotics shall be supported by a business case that demonstrates it improves competitiveness and is not just 'AI for the sake of AI'. |
| Human impact | AI systems should be used to enhance positive social change and enhance sustainability. | 3 | For any deployments, it shall be clear where the human boundary/interface/interaction is with the AI/Analytics/ Robotics system; and any negative/positive impact on human factors and/or human behaviours is fully understood and mitigated where necessary. |
| | | 4 | Early analysis, in conjunction with human resources and employees (or their representatives), shall be undertaken to identify potential job role changes or potential human resource impacts and the opportunities for retraining or redeployment. |
| | | 5 | Potential for upskilling opportunities or redeployment shall be explored with human resources and employees (or their representatives) when any impact on affected employees is established, to ensure that the organisation has the key capabilities neewded to secure emerging opportunities in AI and robotics. |

# Social Impact



| CONTEXT | ETHIC | REALISATION PRINCIPLES | |
|---|---|---|---|
| Human impact | AI systems should be used to enhance positive social change and enhance sustainability. | 6 | Analysis shall be undertaken to assess the impact of the deployment on the supply chain – particularly assessing the likelihood for the technology to have a negative impact on the sustainability of any elements of the supply chain. The same assessment should be afforded to customers as appropriate. |
| | | 7 | Where there is potential for negative impact on the sustainability of the supply chain, this shall be discussed with the supply chain partner as soon as possible to give them maximum opportunity to adapt to remain sustainable. This same opportunity should be afforded to customers as appropriate. |
| Communication | Provide knowledge of the human interactions with AI to key stakeholders. | 8 | Frequent communication and discussion should be had with all key stakeholders – in particular employees and employee representatives – through a variety of channels. |
| Loss of skills | AI systems should be used to enhance positive social change and enhance sustainability. | 9 | Analysis shall be undertaken as to whether any loss/ reduction of skills (which result/cannot be avoided) needs to be sustained, for the good of the business, and how this would be addressed. |

# Accuracy / Trust



| CONTEXT | ETHIC | REALISATION PRINCIPLES | |
|---|---|---|---|
| Safety/zero harm | AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible. | 10 | A formal risk analysis shall be undertaken with specific emphasis given to identifying and mitigating any hazards to human safety. |
| Transparency and Traceability | AI systems must provide for transparency and traceability of their design, inputs and outputs. | 11 | The algorithms shall be assessed for any bias or discrimination impact and their provenance shall be clearly stated to enable any future Root Cause Analysis or troubleshooting (Note, for complex systems, it may be difficult to assess the risk of bias. A new bias assessment template has been created as part of an ecosystem of AI Assurance tools at [link]). |
| | | 12 | To enable the power of data to be unlocked, all training data shall be good quality and representative and its provenance shall be clearly stated to enable any future Root Cause Analysis or troubleshooting. |
| | | 13 | The hierarchy of decision making shall be clearly stated regarding human v AI. |
| | | 14 | It shall be clear what the insight (forecast/decision making etc.) improvement is compared with a human – forecast improvement and actual. |

# Accuracy / Trust



| CONTEXT | ETHIC | | REALISATION PRINCIPLES |
|---|---|---|---|
| Bias | AI systems must be free from bias or prejudice. | 15 | It shall be clearly stated how any training data sets have been assured to have no unintentional or unethical biases, noting that, for example, if an AI sub-system is being used to detect anomalies, the training set may need a deliberate bias to ensure sufficient amounts of anomalies occur at different rates. |
| Validity and Reliability | For AI to succeed it must be trusted. | 16 | A monitor shall be deployed in the system – this is a sense check of the results comparing actual outputs with likely output ranges for the system in question. |
| | | 17 | A continuous automated monitor shall be deployed in the system to continuously test the system by using existing test/synthesised data, which already has known and approved outputs. |
| | | 18 | An independent check shall be deployed in the system – assessment of a proportion of the same data using a completely independent assessment mechanism which is already approved. This is a validation check and could be carried out by a human. |

# Accuracy / Trust



| CONTEXT | ETHIC | REALISATION PRINCIPLES | |
|---|---|---|---|
| Validity and Reliability | For AI to succeed it must be trusted. | 19 | A process comprehensiveness check shall be deployed in the system – ensuring that the right number of assessments have taken place. |
| | | 20 | A faultless transmission of data shall be assured. Where appropriate, a technique such as Cyclic Redundancy Check/checksum should be considered. |
| Sparse data interpolation | For AI to succeed it must be trusted. | 21 | The sparseness of the training set of data and its impact on the validity of the output needs to be clearly stated and justified. |

# Governance



| CONTEXT | ETHIC | REALISATION PRINCIPLES | |
|---|---|---|---|
| Data protection | For AI to succeed it must be trusted. | 22 | It shall be stated whether there is, or will be, any Personal data or not. |
| | | 23 | The legitimate purpose for using the Personal data shall be declared and confirmation provided that this has been agreed with the person or employee representative where it refers to an employee. |
| | | 24 | The architecture of the system shall protect the data from unwanted access without permission - complying with the principle of 'privacy by design and by default'. |
| | | 25 | The architecture of any data storage system should have the facility to, on demand, identify an individual's personal data and update, amend or remove every trace in line with privacy requirements and individuals' rights. |
| | | 26 | No Personal data shall be sent outside of the relevant, legal zone (e.g. European Economic Area, US). |
| Export Control | For AI to succeed it must be trusted. | 27 | The data flows (including access/reading of data) shall be described to, discussed with and approved by an Export Control manager to assure compliance with Export Control regulations. |

# Governance



| CONTEXT | ETHIC | REALISATION PRINCIPLES | |
|---|---|---|---|
| Confidential information | For AI to succeed it must be trusted. | 28 | All confidential information shall be declared to, discussed with and the architectural protections approved by an IT security expert. |
| Cyber security | For AI to succeed it must be trusted. | 29 | All confidential information shall be declared to, discussed with and the architectural protections approved by an IT security expert. |
| Accountability | Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. | 30 | Ultimate accountability for the outcomes of the AI system needs to be clearly stated with a business owner clearly identified. |
| Responsibility for decisions | Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. | 31 | Algorithmic accountability should fall jointly on the developer and tester, or the DevOps team. They shall clearly state how they have assured confidence in the performance of their individual aspects of the AI system. |
| Risks from re-use/transfer across processes | For AI to succeed it must be trusted | 32 | Transferring knowledge between AI systems should be risk assessed using a formal tool/method to determine where and how the system might fail. Any serious events and their causes must be identified along with the method to detect such events. – which shall be formally reviewed before proceeding. |