**What is responsible Artificial Intelligence (AI)?**

Artificial intelligence is a science and technology based on disciplines such as Computer Science, Biology, Psychology, Linguistics, Mathematics, and Engineering. A major thrust of AI is in the development of computer functions associated with human intelligence, such as reasoning, learning, and problem solving.

Responsible AI is about creating governance frameworks to evaluate, deploy and monitor AI to create new opportunities for better citizen and mission services. It means building and implementing solutions that put people at the centre

There are 4 key areas that have been identified for responsible AI practices. These are:

* governance,
* ethically design solutions,
* risk control and training and
* education

**Instances where AI has failed or been used maliciously or incorrectly**

***Microsoft chatbot goes Nazi on Twitter:***

Back in 2016, Microsoft launched a chatbot (called ‘Tay’) which it said was designed to engage and entertain people through casual and playful conversation. They said, that Tay gets smarter the more that you chat with her. However, Microsoft ran into a PR nightmare when Tay went radically off message and started to spout abusive statements and even Nazi sentiments. It Tweeted things like “Hitler was right’” and “9/11 was an inside job”. Tay was parroting offensive statements made by other humans trying to provoke her. Tay was designed to mimic the language patterns of millennial females in the 18 - 24-year-old demographic. By way of machine learning and algorithms, Tay could approximate conversation by processing inputted phrases and blending in other relevant data. Essentially, Tay found herself mixing with the wrong people, and was taken offline after 16 hours.

***Amazon – failure of its facial Recognition system – Rekognition***

Amazon launched a facial recognition system which was marketed to police agencies for use in their investigations. The system was tested by the Massachusetts American Civil Liberties Union, which found that the database incorrectly matched the images of professional athletes to a database of mugshots and falsely identified some as criminals. Nearly 1 in 6 athletes were falsely identified.

The misclassification was am embarrassment to Amazon.

**Other Failures include:**

* The self-driving Uber, which failed to stop at a red signal at a busy crossing in San Francisco
* Amazons Alexa was promoted to order items for hundreds of households after being triggered by an advert on the TV
* Amazon – tested a recruitment tool which was filtering applicants by gender based upon key words it was trained to look for.

**Why AI fails?**

There could be different reasons for this, but most of them can be grouped as follows:

Bad or insufficient data

The data is the most important component of AI now. Training data is used to train statistical models by means of machine learning (deep learning) algorithms. Typically, a deep learning task would require millions of data items to build an adequate model. The data must be a good reflection of the real-world situation, without any bias that may lead to the situation when one class is preferred to another without much reason, just because the data had more samples of one class. This may be the case with the Amazon recruitment AI failure.

Bad engineering

It’s always difficult to spot, if a particular failure is due to bad engineering, because it is usually impossible to go through the code and get all the neural network settings to be sure that the problem is there (e.g. but not in the data). Of course, in the examples described here we can safely assume that engineering was good, because the companies mentioned here can afford the best of the best engineers.

Wrong area of application

It may happen that the data used for some task is too complex, difficult to obtain, or that the result needs to be much more accurate, than any algorithm can achieve. For example, application of AI techniques to healthcare, law enforcement, and other industries may be too risky. Even in hospitality industry robots may become a problem rather than solution, because people need more than just to be served.

**How to avoid such failures?**

Never overestimate the capabilities of AI. It doesn’t make miracles and it is nowhere close to those “strong AI” smarties from Hollywood blockbusters. You need a lot of relevant, cleaned and verified data to train an adequate model. The data is crucial for machine learning, but it is not all you need. Choosing a correct algorithm and tuning its parameters need a lot of tests and trials by a team of highly qualified experts.

Most importantly, an AI system has a very limited capability of replacing humans. It can replace humans in simple, but tedious tasks, that consist of a lot of repeating routines. Any complex task that requires non-trivial approach to solution may lead to a high level of errors by AI. The best role, an AI can play now is an assistant to humans who use AI as a tool to do a lot of routines and repeating operations.

**AI and GDPR Law – what does it cover, especially with regards to automated decision making.**

GDPR’s Article 22 [covers](https://gdpr-info.eu/art-22-gdpr/) “automated individual decision-making, including profiling.” Some scholars [assert](https://iapp.org/news/a/want-europe-to-have-the-best-ai-reform-the-gdpr/) that this provision could lead AI companies to limit activities such as offering customers loans or to implement additional and expensive human review of AI-powered decisions. Others, such as [Kalliopi Spyridaki](https://iapp.org/about/person/0011a00000DlNa2AAF/), chief privacy strategist at SAS Institute Inc., [argue](https://www.sas.com/en_us/insights/articles/data-management/gdpr-and-ai--friends--foes-or-something-in-between-.html) that although the GDPR may sometimes limit or complicate how AI technologies use data, the GDPR could also “help create the trust that is necessary for AI acceptance by consumers and governments.

* At a glance, the GDPR has provisions on:
  + automated individual decision-making (making a decision solely by automated means without any human involvement); and
  + profiling (automated processing of personal data to evaluate certain things about an individual). Profiling can be part of an automated decision-making process.
* The GDPR applies to all automated individual decision-making and profiling.
* Article 22 of the GDPR has additional rules to protect individuals if you are carrying out solely automated decision-making that has legal or similarly significant effects on them.
* You can only carry out this type of decision-making where the decision is:
  + necessary for the entry into or performance of a contract; or
  + authorised by Union or Member state law applicable to the controller; or
  + based on the individual’s explicit consent.
* You must identify whether any of your processing falls under Article 22 and, if so, make sure that you:
  + give individuals information about the processing;
  + introduce simple ways for them to request human intervention or challenge a decision;
  + carry out regular checks to make sure that your systems are working as intended.

The text of Article 22 GDPR is copied below:

# Art. 22 GDPR

# Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
   1. is necessary for entering into, or performance of, a contract between the data subject and a data controller;
   2. is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests; or
   3. is based on the data subject’s explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in [Article 9](https://gdpr-info.eu/art-9-gdpr/)(1), unless point (a) or (g) of [Article 9](https://gdpr-info.eu/art-9-gdpr/)(2) applies and suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests are in place.

**What should organisations do to ensure they are being responsible with AI and the wider use of data in general.**

For organisations to ensure they are being responsible with AI, they need to make sure that their use of AI fulfils a number is criteria. These are firstly that the AI is ethically sound and complies with regulations in all respects; secondly, the AI is underpinned by a robust foundation of end-to-end governance; and thirdly, the AI is supported by strong performance pillars addressing bias and fairness, interpretability and explain-ability and robustness and security.

A Comprehensive Responsible AI Framework needs to address 5 key dimensions:

* 1. **GOVERNANCE**

This serves as an end-to-end foundation for all the other dimensions. At its highest level, AI governance should enable an organisation to answer critical questions about results and decision-making of AI applications, including:

• Who is accountable?

• How does AI align with the business strategy?

• What processes could be modified to improve the outputs?

• What controls need to be in place to track performance and pinpoint problems?

• Are the results consistent and reproducible?

* 1. **ETHICS & REGULATION**

The core goal here is to help organisations develop AI that is not only complaint with applicable regulation bit is also ethical. Organisations should strive to develop, implement, and use AI solutions that are both morally responsible and also legal and ethically defensible. By contextualising the ethical considerations, organisations may be able to identify the ethical implications of their AI solutions, and the relevant principles that should be taken into account when designing and operationalising AI models, allowing for robust mitigation of ethical risks.

* 1. **INTERPRETABILITY & EXPLAINABILITY**

This provides an approach and utilities for AI-driven decisions to be interpretable and easily explainable by those who operate then and those who are affected by them. A lack of interpretability in AI decisions is not only frustrating for end-users or customers, but can also expose an organisation to operational, reputational, and financial risks. To instil trust in AI systems, people must be enabled to look “under the hood” at their underlying models, explore the data used to train them, expose the reasoning behind each decision, and provide coherent explanations to all stakeholders in a timely manner.

* 1. **ROBUSTNESS & SECURITY**

Helps organisations develop AI systems that provide robust performance and are safe to use by minimising the negative impact.

To be effective and reliable, AI systems need to be resilient, secure, and safe. In terms of resilience, next-generation AI systems are likely to be increasingly “self-aware,” with a built-in ability to detect and correct faults and inaccurate or unethical decisions. In terms of security, the potentially catastrophic outcomes of AI data or systems being compromised or “hijacked” make it imperative to build security into the AI development process from the start, being sure to cover all AI systems, data, and communications.

Above all, though, AI systems must be safe for the people whose lives they affect, whether they are users of AI or the subjects of AI-enabled decisions. This is clearly crucial in areas such as healthcare, autonomous vehicles, and connected worker or manufacturing applications.

* 1. **BIAS & FAIRNESS**

Addresses the issues of bias and fairness—recognising that while there is no such thing as a decision that is fair to all parties, it is possible for organisations to design AI systems to mitigate unwanted bias and achieve decisions that are fair under a specific and clearly communicated definition.

That said, even having clear criteria doesn’t remove the need to make trade-offs between competing—and sometimes conflicting—priorities. In this sense, AI decisions are similar to those made by humans. In each case, establishing fairness requires businesses to choose their level of comfort in the choices they make, and balance these against the associated costs and wider impacts, which might be negative for some.

**EXTENSION TASK**

**An article from UK Tech News**

<https://uktechnews.co.uk/2020/04/01/what-transformational-business-leaders-need-to-know-about-artificial-intelligence/>

This article will examine the key pillars of an effective AI strategy and outline the challenges that it will need to address––talent, time and trust –– and the best practices to ensure AI success in the long-term, and what business and IT leaders need to know to achieve a successful AI Transformation across their organisation.

## 1: Talent: Data is a Team Sport

Because AI can drive better business results, there are fundamental aspects of an organisation that must be considered. Ultimately, the long-term success of an AI strategy will depend on the people and culture within the company.

Firstly, the current technical skills gap and lack of AI talent available must be addressed. Research from EY indicates that 56 percent of tech professionals involved with AI say that the lack of available talent is the biggest barrier to increasing AI adoption amongst businesses. As a role that is in extremely high demand, expert level data scientists are being quickly snapped up by tech companies and the internet giants as soon as they hit the job market, leaving many companies with a much smaller talent pool to hire from.

Simply put, there are not enough AI experts available in the workplace. More people need to pursue a career in data science and AI, and this can be accomplished by the implementation of STEM programmes at the primary, middle school and university levels. In addition, machine learning should be fast, accurate and available to everyone. By democratising AI, we can further advance our understanding and awareness of data and decision-making for all businesses. To achieve this, businesses should find a technology that can simplify machine learning and data science problems, even if an organisation doesn’t have a dedicated data science expert on its team.

However, when a company does have data scientists, they must understand that data is a team sport. Getting people with different skill sets to work together effectively, enabling teamwork across an organisation and working well as a team to make the data work for them is crucial to building a successful data-driven business. The “data team” consists of everyone from the functional business leaders to devops professionals and analysts, to data engineers and data scientists. Culturally, this team must be collaborative in order to be transformative. They must learn to work within the existing culture of a company, to bring a lasting positive change.

## 2: Time: Obtaining Results Faster

Using data is a great way to make informed decisions. But how do you glean insights from data that enables more efficient and effective decisions? Business leaders are generally overwhelmed with data from multiple areas of their organisation and need to address a range of use cases that are primed for AI, including how to attract new customers, make a credit-scoring decision, detect fraud or pinpoint the correct treatment for a certain patient.

Essentially what many businesses are trying to do is extract real insights from data. To make the best and most informed decisions requires not just data, but also time. AI can assist in making the right decisions, easier, cheaper and in record time. By building AI models, data science teams can visualise every scenario based on data that the company already has. Data can then be used to re-train the model in the future, allowing it to continuously learn and improve. IT and business leaders seek out solutions that can help speed time to insights and time to better results.

## 3: Trust: Explain the AI

Debatably, the biggest obstacle preventing AI success is trust. As organisations assemble strong data and AI teams, trust in the technology itself is one of the most important ingredients to the successful integration of AI into a company’s culture and business processes. For instance, how will people within an organisation trust an algorithm over the decades of existing human intuition and experience?

To overcome this challenge requires a more comprehensive explanation of AI to people within the organisation about what the technology is, how it will be utilised and how it will ultimately help people by allowing them to accomplish more, faster, in order to complete larger, more creative projects, and solve more critical and complex problems. The goal here is to have AI running in the background at all times – it’s the permanent Plan B. Plan A is still to use your manual tool base, namely the humans who work at the company. New technologies on the market today can address the explainability and interpretability of a model, so also factor that in when considering a solution.

## 4: Influencing Change via a Maker Culture

From an organisational and cultural perspective, it’s crucial to remember that AI is not simply a new fad, but it is the catalyst for a chain reaction in the direction of change. With AI, organisations can create a maker culture, where learning is best done through doing. AI is capable of instilling a product culture that continues its life cycle inside the businesses processes. As companies adopt AI strategies, they become makers and influence change.

It has been both fascinating and exciting to witness the development of AI over the last few years, as companies begin adopting AI strategies. It will only get more interesting and rewarding for businesses moving forward, as they continue on their AI Transformation journey.