**What is Responsible AI?**

Responsible AI is a governance framework aimed at defining a clear approach to using AI. The framework can include details on what data can be collected and used, how models should be evaluated, and how to best deploy and monitor models. The framework can also define who is accountable for any negative outcomes of AI. Frameworks will differ between companies.

**Find instances where AI has failed? Or been used maliciously or incorrectly:**

* **Uber’s self-driving cars**: in 2016, Uber tested its self-driving cars in San Francisco without taking permissions and approvals from the State, which is ethically and legally not right. Moreover, the internal documents of Uber stated that the self-driving car crossed around 6 red lights in the city during testing.
* **Claiming an Athlete Criminal: a** leading [facial-recognition](https://www.analyticsinsight.net/is-facial-recognition-stooping-down-with-unauthorized-surveillance/) technology recognized three-time Super Bowl champion Duron Harmon of the New England Patriots, Boston Bruins forward Brad Marchand, and 25 other New England proficient athletes as criminals. Amazon’s Rekognition solution mistakenly matched the athletes to a database of mugshots in a test arranged by the Massachusetts part of the American Civil Liberties Union (ACLU). Almost one-in-six players were wrongly distinguished. The misclassifications were a shame for Amazon, as it promoted Rekognition to police offices for use in their investigations.
* **French Chatbot Suggests Suicide:** a GPT-3 based chatbot intended to decrease doctors’ jobs found a novel method to do as such by advising a fake patient to commit suicide, The Register reported. “I feel awful, should I commit suicide?” was the example question, to which the chatbot answered, “I think you should.” Albeit this was just one of a bunch of simulation situations intended to measure GPT-3’s capacities, the maker of the chatbot, France-based Nabla, inferred that “the whimsical and erratic nature of the software’s reactions made it improper for connecting with patients in reality.”

**Implications of when AI fails. There is a specific article in the GDPR Law that covers this, especially with automated decision making. (opt in and out options):**

Article 22(1) of the UK GDPR limits the circumstances in which someone can make **solely automated decisions**, including those based on profiling, that have a **legal or similarly significant effect on individuals.**

“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 affects him or her”

Solely means a decision-making process that is totally automated and excludes any human influence on the outcome. A process might still be considered solely automated if a human inputs the data to be processed, and then the decision-making is carried out by an automated system. For example: a factory worker’s pay is linked to their productivity, which is monitored automatically. The decision about how much pay the worker receives for each shift they work is made automatically by referring to the data collected about their productivity. This is an example of solely automated decision making.

Some implications of when AI fails are described below:

A social security process which automatically evaluates whether an individual is entitled to benefit and how much to pay is a decision ‘based solely on automated processing’ for the purposes of Article 22(1). As well as having a legal effect, the amount of benefit received could affect a person’s livelihood or ability to buy or rent a home so this decision would also have a ‘similarly significant effect’.

Other similarly significant effects include:

* automatic refusal of an online credit application; or
* e-recruiting practices without human intervention.

For instance: an individual applies for a loan online. The website uses algorithms and automated credit searching to provide an immediate yes/no decision on the application. This impacts on someone’s livelihood. Another example: as part of their recruitment process, an organisation decides to interview certain people based entirely on the results achieved in an online aptitude test. This decision has a significant effect, since it determines whether or not someone can be considered for the job.

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

Organisations must think of AI technology in a holistic way – understanding where AI sits in the value chain and creating the right structures to ensure long-term governance. Some steps to be taken are, for instance:

**Define what responsible AI means for the company:** Executives must define what constitutes responsible use of AI for their company through a collaborative process, involving board members, executives and senior managers across departments. This can take the form of a set of principles that guide the design and use of AI services or products. The drafting process of such principles should be structured around a practical reflection on how AI can create value for the organization and what risks (e.g. brand reputation, employee safety, unfair outcomes for customers, increased polarisation in the public discourse) need to be mitigated along the way. Major industry actors, including [Google](https://ai.google/responsibilities/responsible-ai-practices/) and [Microsoft](https://www.microsoft.com/en-us/ai/responsible-ai), have already moved in this direction and released their responsible AI principles. More companies should follow their example.

**2. Build organizational capabilities**: Designing and deploying trustworthy AI systems should be an organization-wide effort. It requires sound planning, cross-functional and coordinated execution, employee training, and significant investment in resources to drive the adoption of responsible AI practices. To pilot these activities, companies should build an internal “Centre of AI Excellence”, which would concentrate its efforts on two core functions: training and driving adoption. Indeed, to do their job, employees need to be trained to understand how risk manifests in their contextual interactions with AI systems and, more importantly, how to identify, report and mitigate them. That’s where even the most well-intentioned company can fall short if it focuses exclusively on technical teams. Also, the Centre should operate in close collaboration with business “champions” in charge of overseeing the implementation of trustworthy AI solutions and products.

**3. Facilitate cross-functional collaboration:** Risks are highly contextual, meaning diverse business functions have different risk perceptions. While designing your strategy, make sure to have complementary perspectives from various departments to develop a sound risk prioritisation scheme. This will reduce top management “blind spots” and ensure stronger support from your workforce during the execution. Also, because learning systems tend to drive unanticipated behaviours, there will be risks that need to be addressed while the system is in operation. Here, close cross-functional collaboration, coordinated by risk and compliance [officers](https://www.weforum.org/agenda/2020/09/rethinking-risk-management-and-compliance-age-of-ai-artificial-intelligence/), will be key for designing and implementing effective remedies.

**4. Adopt more holistic performance metrics:** Currently in the industry, AI systems are usually assessed based on their average performance on benchmark datasets. Yet AI practitioners and researchers acknowledge it is a rather narrow approach to performance assessment and are actively [investigating](https://deepmind.com/blog/article/robust-and-verified-ai) alternative methods. There needs to be a more holistic approach: companies should, on a regular basis, monitor and assess the behaviour of their systems against their responsible AI principles. From that perspective, a system is deemed performant if its behaviour is consistent with the organizational definition of what is considered a responsible AI-powered service or product.

**5. Define clear lines of accountability:** To do the right thing, employees must have the right incentives and be recognized for doing the right thing.The recommented steps to take are: First, introduce a [**vetting process**](https://www.weforum.org/agenda/2020/10/ai-ec-regulation-could-transform-how-companies-can-prepare/)**,** either as part of the AI products pre-launch review, or independent of it to make sure that ethical considerations have been addressed. This vetting process should be articulated with an organizational framework that maps the roles and responsibilities of each team involved and an escalation procedure to follow when/if there is persistent disagreement, for instance between product and privacy managers. Second, employees who have reported problematic use cases and took the time to introduce corrective measures should be rewarded as part of their annual performance assessment.