|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Artificial Intelligence (AI) HLT  [Document subtitle] |
| |  |  |  | | --- | --- | --- | | Fiona Warigon |  | Data Analysis | |

# What Is Responsible AI

Responsible AI is the practice of designing, developing, and deploying AI with good intention to empower employees and businesses, and fairly impact customers and society—allowing companies to engender trust and scale AI with confidence.

# Instances Where AI Has Failed

AI has failed in the following ways.

**Automation of social engineering attacks:** Victims’ online information is used to automatically generate custom malicious websites/emails/links they would be likely to click on, sent from addresses that impersonate their real contacts, using a writing style that mimics those contacts.

**Automation of vulnerability discovery:** Historical patterns of code vulnerabilities are used to speed up the discovery of new vulnerabilities, and the creation of code for exploiting them.

**More sophisticated automation of hacking:** AI is used (autonomously or in concert with humans) to improve target selection and prioritization, evade detection, and creatively respond to changes in the target’s behaviour. Autonomous software has been able to exploit vulnerabilities in systems for a long time, but more sophisticated AI hacking tools may exhibit much better performance both compared to what has historically been possible and, ultimately (though perhaps not for some time), compared to humans.

**Human-like denial-of-service:** Imitating human-like behaviour (e.g., through human-speed click patterns and website navigation), a massive crowd of autonomous agents overwhelms an online service, preventing access from legitimate users and potentially driving the target system into a less secure state.

**Automation of service tasks in criminal cyber-offense:** Cybercriminals use AI techniques to automate various tasks that make up their attack pipeline, such as payment processing or dialogue with ransomware victims. Prioritising targets for cyber-attacks using machine learning. Large datasets are used to identify victims more efficiently, e.g., by estimating personal wealth and willingness to pay based on online behaviour.

**Terrorist repurposing of commercial AI systems:** Commercial systems are used in harmful and unintended ways, such as using drones or autonomous vehicles to deliver explosives and cause crashes.

# Implications Of When AI Fails

The following are implications when AI fails:

* AI Bias.
* Loss of Certain Jobs.
* Global Regulations.
* Accelerated Hacking.
* AI Terrorism.

# What Should Organisations Do to Ensure That They Are Being Responsible With AI

 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 by:

1. Establishing internal governance, for example by an objective review panel, which is diverse and that has the knowledge to understand the possible consequences of AI infused systems. A key success factor is leadership support and the power to hold leadership accountable.
2. Ensuring the right technical guardrails, creating quality assurance and governance to create traceability and auditability for AI systems. This is an important part of every organisation’s toolkit to allow operational and responsible AI to scale.
3. Investing more in their own AI education and training so that all stakeholders – both internal and external – are informed of AI capabilities as well as the pitfalls.