Core Learning 1 –Artificial Intelligence

# What is responsible AI?

Responsible AI is a method of developing and using AI to be as safe, trustworthy, and ethical as possible. Microsoft’s ‘responsible AI standard’ uses six principles of fairness, reliability/safety, privacy/security, inclusiveness, transparency and accountability. (Microsoft, 2022)

Fairness and inclusiveness mean considering similar groups of people in an unbiased way - for example, treating similarly qualified applicants for a job the same.

Reliability/safety means responding to unanticipated situations safely, resisting harmful manipulation, and operating reliably and consistently as designed.

Transparency means allowing stakeholders to understand why AI behaves the way it does, in order to identify potential performance or fairness issues.

Privacy/security means AI complying with data privacy and consumer control of data laws.

Accountability means AI designers and users should be held accountable for AI systems based on industry standards to ensure that AI systems don’t have final authority over decisions that can impact the lives of people.

# Where has AI failed, been used maliciously or been used incorrectly?

One example of AI failing is the Twitter chatbot created by Microsoft, Tay, which was released in 2016. It was designed to interact with young adults on the platform, but because it learns to communicate in part based on interactions with Twitter users, it quickly learned to swear and make offensive remarks. Microsoft shut it down only 16 hours after launch. (The Guardian, 2016)

Another example is Neuro-sama, an AI Twitch streamer who plays games and interacts with the audience. Released in December 2022, Neuro-sama responds to questions that the audience asks her, and when she was asked if she had heard of the Holocaust, she replied “I’m not sure I believe it.”, which resulted in Twitch banning the channel from streaming for two weeks. The creator, Vedal, responded by implementing measures to automatically filter offensive content, and manually cancel statements that might violate Twitch’s terms of service. (Wikipedia, 2023)

A third example is *Nothing, Forever*, a procedurally generated sitcom based on Seinfeld, which, like Neuro-sama, was also streamed from December 2022 on Twitch and was also banned for two weeks for making an offensive remark, although in this case the remark was not caused by audience input, but by the creators switching to a less sophisticated GPT model due to outages. The creators responded by fixing the issues they had with the more sophisticated model and switching back to it. (Engadget, 2023)

# What are the implications of AI failure?

AI failure can have disastrous consequences, depending on what application the AI is being used in. In 2018, Elaine Herzberg became the first person to be killed by a self-driving car when she was hit by a Volvo being tested by Uber in Tempe, Arizona, despite the presence of a backup driver. According to investigators, the backup driver was watching television at the time. This highlights the importance of exercising strict control over experimental AI to ensure that people in the presence of such technologies are as safe as possible. It also raises the issue of accountability, as Uber was not held criminally responsible, although they were accused by regulators as having an “inadequate safety culture”, and days before the crash, an employee had warned that the vehicles were unsafe and the backup drivers poorly trained. (BBC News, 2020)

The examples cited in Question 2 show how AI can negatively impact fairness and inclusiveness, as some remarks can be offensive to certain groups and make them feel excluded.

The Data Protection Act 2018 is the government’s implementation of the EU regulation ‘General Data Protection Regulation (GDPR)’ and ensures that personal data is used fairly, lawfully and transparently. One aspect of the law ensures that you have rights when an organisation is using your data for automated decision-making processes. (Gov.uk, n.d.). This relates to Article 22 of the GDPR, which states that people have a 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.”, without the person’s explicit consent. (GDPR-Info.eu, n.d.). One example of how personal data could be used to create a biased automated decision is gender, racial or other discrimination in calculating insurance premiums.

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

One thing organisations could do is ensure AI is built and used in accordance with the principles of fairness, reliability/safety, privacy/security, inclusiveness, transparency and accountability. Another thing they could do is to ensure well-trained human controllers have the final say in what an AI is going to do in any situation where there is a risk of harm, especially if the AI is still in its early stages of development.

Organisations should establish an objective review panel to understand the consequences of systems which use AI. Technical safeguards should be put in place to ensure that AI systems are traceable and auditable, creating quality assurance. Finally, education should be prioritised for all stakeholders so that the risks of AI are better understood. (Rosenshine, 2020)

# References

BBC News. (2020). Retrieved from https://www.bbc.co.uk/news/technology-54175359

Engadget. (2023). Retrieved from https://www.engadget.com/the-ai-seinfeld-show-nothing-forever-is-back-on-twitch-064359854.html

GDPR-Info.eu. (n.d.). Retrieved from https://gdpr-info.eu/art-22-gdpr/

Gov.uk. (n.d.). Retrieved from https://www.gov.uk/data-protection#:~:text=The%20Data%20Protection%20Act%202018%20is%20the%20UK's%20implementation%20of,used%20fairly%2C%20lawfully%20and%20transparently

Microsoft. (2022). Retrieved from https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai

Rosenshine, K. (2020). Retrieved from Microsoft: https://www.microsoft.com/en-gb/industry/blog/cross-industry/2020/01/08/3-ways-organisations-can-use-ai-in-a-responsible-way/

The Guardian. (2016). Retrieved from https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot

Wikipedia. (2023). Retrieved from https://en.wikipedia.org/wiki/Neuro-sama