Week 7 HLT- Responsible AI

What is Responsible AI?

Responsible AI is a framework containing a set of principles that must be applied to AI systems to ensure that they are interpretable, fair, safe and respectful of a user’s privacy. These principles cement a user’s trust in an AI system.

Instances where AI has failed

On March 18th 2018, the death of Elaine Herzberg was the first recorded case of a pedestrian fatality involving a self-driving car with a safety driver present (but who didn’t take sufficient action to stop the vehicle). This fatal error was caused by the car’s training that labelled objects as pedestrians if “that object was near a crosswalk”.

Another instance of an AI failure was when a soap dispenser would not dispense soap for black people. This was caused by how the dispenser worked. The soap dispenser emitted infrared radiation and if the radiation is reflected (reflected off a hand), then the AI interprets this as ‘soap must be dispensed’. However, infrared radiation is not reflected by black objects and reflected by white objects. So black hands reflect much less infrared radiation than white hands which caused the problem. This highlighted a problem surrounding lack of diversity in tech companies as this approach to the solution (infrared radiation) would not have been suggested if there was a more diverse team implementing this AI.

Implications of when AI fails

In addition to losing user’s trust in AI systems in general, when AI fails, it highlights a company’s negligence to that particular issue. The implication of this is that there is usually legal action against the company that implemented the AI in question. For example, in the self-driving car fatality described above, there was legal action against Uber (the company that implemented the AI) and an undisclosed settlement was reached for Herzberg’s daughter.

However, AI is an emerging field and laws and regulations must be kept up to date with the pace of AI development. This means external bodies must update and uphold regulations for AI

systems. For instance, Article 22 in the GDPR regulations states that users have the right for their personal data to not be used by a fully automated process if the process involves making predictions about the individual (profiling). The user can opt in to this process if they wish.

What should organisations do to ensure that they are being responsible with AI?

Organisations should be transparent about how their AI makes predictions and how user’s data is used in their AI systems. For instance, organisations can provide users with pre-made explanations for how the AI system makes predictions if requested. This is used as a check to ensure that users are not discriminated against. Additionally, organisations should inform users how their data will be used and for how long the data will be stored (in compliance with GDPR regulations).

Organisations must ensure their AI systems are secure and encrypted to protect against malicious manipulations of the systems, which can lead to their AI making incorrect decisions.

Sources

[What is Responsible AI? | by Conor O'Sullivan | Towards Data Science](https://towardsdatascience.com/what-is-responsible-ai-548743369729)

<https://towardsdatascience.com/5-ai-failures-you-probably-should-know-about-417ddebbc323>

<https://en.wikipedia.org/wiki/Death_of_Elaine_Herzberg>

<https://www.standard.co.uk/news/world/automatic-soap-dispenser-sparks-racism-outrage-after-footage-shows-it-doesn-t-work-for-darkskinned-people-a3615096.html>

<https://gdpr-text.com/read/article-22/#comment_gdpr-a-22_1>