Case Study 5

AI-Based Cyberbullying Interventions – Evaluating Youth Perspectives

Theme of the Case Study:

- Online Safety
- Bias in AI Moderation
- Youth Digital Wellbeing

Abstract (Summary):
This study investigated youth perceptions of AI-driven cyberbullying interventions on social media. Researchers at Dublin City University (DCU) tested AI-based proactive content moderation strategies designed to detect and limit harmful interactions online. The consultation involved young people aged 12 to 17, who evaluated these interventions through focus groups and online discussions.

The study found that while AI content moderation can reduce harmful interactions, youth participants expressed concerns over:
• False positives and censorship – AI incorrectly flagging harmless content as harmful.
• Lack of human context – AI struggling to understand nuances in humour and informal language.
• Privacy concerns – Uncertainty about who controls AI interventions and their impact on digital rights.

The research emphasised the need for youth participation in AI design, ensuring that content moderation aligns with young people's perspectives and digital realities.

Relevance to the Reader:

Many young people rely on AI-driven platforms for social interaction and education, making ethical AI implementation critical.

Cyberbullying is a growing concern, and while AI moderation can help, it must be transparent, fair, and accountable to users. This case study encourages youth workers and educators to discuss AI's role in digital safety, fostering critical thinking on algorithmic decision-making. • The study highlights the importance of engaging young people in the design and evaluation of AI interventions, ensuring that solutions are youth-centred and ethically sound.