Online Safety 4 Schools

Online Safety 4 Schools

Safeguarding in the Age of Artificial Intelligence

Artificial Intelligence (AI) is transforming education, healthcare, social care, business, and everyday life. While AI creates opportunities for innovation, efficiency, and personalised support, it also introduces new safeguarding challenges that organisations, professionals, and communities must address. Effective safeguarding in the age of AI requires balancing technological advancement with the protection of individuals, particularly children and vulnerable adults.

What is AI Safeguarding?

AI safeguarding refers to the policies, practices, and measures used to protect people from harm arising from the development, deployment, and use of artificial intelligence. This includes preventing:

- Online exploitation and grooming
- Cybercrime and fraud
- Privacy breaches and misuse of personal data
- Deepfakes and misinformation
- Bias and discrimination in automated decision-making
- Psychological and emotional harm
- Unsafe or inappropriate AI-generated content
- Overreliance on AI in critical decisions

AI presents both opportunities and risks, and safeguarding frameworks must evolve to address these emerging challenges.

Key Safeguarding Risks

1. Deepfakes and AI-Generated Abuse

Generative AI can create realistic images, videos, and audio recordings that appear authentic. These technologies can be misused to:

- Create non-consensual intimate images
- Facilitate cyberbullying and harassment
- Spread false information
- Damage reputations

For children and young people, deepfakes present significant safeguarding concerns because victims may experience lasting emotional and psychological harm.

2. Online Grooming and Exploitation

AI tools can generate realistic profiles, messages, and conversations that enable offenders to deceive young people online. Criminals may use AI-generated identities to build trust, manipulate victims, and facilitate exploitation or sextortion.

3. Data Privacy and Protection

Many AI systems rely on large amounts of user data. Risks include:

- Unauthorised data collection
- Data breaches
- Profiling and surveillance
- Misuse of sensitive information

Organisations must ensure compliance with data protection legislation and carefully assess how AI systems collect, store, and process personal information.

4. Bias and Discrimination

AI systems learn from historical data, which may contain societal biases. Without careful oversight, AI can:

- Reinforce inequalities
- Produce discriminatory outcomes
- Disadvantage vulnerable groups

Safeguarding therefore includes ensuring fairness, transparency, and accountability in AI-driven decisions.

5. Misinformation and Manipulation

AI can rapidly generate convincing text, images, and videos. This increases the risk of:

- Fake news
- Health misinformation
- Political manipulation
- Financial scams

The ability to distinguish genuine information from AI-generated content is becoming a critical digital literacy skill.

6. Cybersecurity Threats

AI systems themselves can become targets for cyberattacks or be used by attackers to increase the sophistication of phishing, fraud, and malware campaigns. Organisations need robust cybersecurity controls throughout the AI lifecycle.

7. Cognitive Dependence and Skill Erosion

Increasing reliance on AI tools may reduce critical thinking, problem-solving, and decision-making skills if individuals become overly dependent on automated systems. Professionals must retain the ability to challenge and verify AI outputs.

Opportunities for Safeguarding

AI is not only a source of risk—it can also strengthen safeguarding efforts.

Examples include:

- Detecting online abuse and harmful content
- Identifying patterns of exploitation
- Supporting early intervention for vulnerable individuals
- Monitoring cyber threats
- Improving accessibility for people with disabilities
- Enhancing educational support and personalised learning

When used responsibly, AI can become a powerful safeguarding tool.

Principles for Safe AI Use

To safeguard effectively in the AI era, organisations should adopt the following principles:

Human Oversight

Critical decisions affecting welfare, safety, healthcare, education, or social care should always involve human judgment.

Transparency

Users should understand when AI is being used and how decisions are made.

Accountability

Clear responsibility must exist for AI-related decisions and outcomes.

Privacy by Design

Data protection should be built into AI systems from the outset.

Fairness and Inclusion

Regular testing should identify and address bias and discriminatory outcomes.

Digital Literacy

Staff, parents, carers, and young people should receive education about AI opportunities and risks.

Continuous Review

AI technologies evolve rapidly; safeguarding policies must be regularly updated.

Conclusion

Safeguarding in the age of AI is no longer a future concern—it is a present responsibility. Organisations must recognise that AI introduces new forms of harm while also offering powerful tools for protection and support. The challenge is not simply to restrict AI, but to ensure that its use is ethical, transparent, secure, and centred on human wellbeing. By combining robust governance, digital literacy, and human oversight, society can harness the benefits of AI while protecting those most vulnerable to its risks.

Jonathan Taylor MSc
Online Safety & Social Media Consultant

Website: www.onlinesafety4schools.co.uk
Email: onlinesafety4schools@gmail.com