**9. Platform Governance and Content Moderation**

**9.a. What is Platform Governance?**

“Governance” refers to the system of rules, policies, practices and standards that manages how an organisation operates and the use of mechanisms to hold organisations and individuals to account.[[1]](#footnote-1) Thus, “platform governance” refers to the management of platforms, such as social media, their users and stakeholders, using a range of rules, policies, practices and standards to govern the use, operation and purpose of the platform.

**9.a.i. Self-Regulation Model of Platform Governance**

There are many different models of platform governance, however, the most common type is the self-regulation model. Many online platforms hold an extreme amount of discretion as to how to govern their platforms and whilst there have been recent improvements in content moderation, self-regulation is still the prominent form of platform governance.[[2]](#footnote-2)

**9.b. Responsibilities of Platforms in Content Management**

Online platforms have a responsibility to manage the content available to its users. Content moderation ensures that user-generated content meets standards and guidelines through monitoring and reviewing.[[3]](#footnote-3) Content moderation is a form of platform governance, and its goal is to enforce community guidelines and terms of service. There are various strategies that platforms use to govern the content including: human moderation and automated moderation.

**9.b.i Human Moderation**

Human content moderation is a form of manual moderation. This typically involves the manual monitoring of user generated content to ensure that the content conforms to the platform’s guidelines. User-generated content that does not conform to the law, guidelines or platform specifications is removed through human intervention to improve user satisfaction and protect users from illegal or harmful content.

Manual human moderation has many benefits including the ability to understand context behind content, which improves content-specific judgements within full-scale datasets. However, its major downfall occurs in its cost and time effectiveness. Human moderation is unable to manage content to the same scale at the same price as automated moderation.[[4]](#footnote-4)

**9.b.ii. Automated Moderation**

Automated moderation includes any form of automated response to user-generated content submitted to a platform. Automated moderation includes tools like algorithms and artificial intelligence (AI) to ensure that content meets platform guidelines. The use of AI in automated moderation allows user-generated content to be reviewed against platform data and determine an appropriate cause of action – the AI itself is able to learn from these outcomes and develops more accurate moderation. The benefits of automated moderation are centred around the time and cost benefits. Automated moderation can make almost instantaneous decisions about content, ensuring it meets regulations. This creates a cheaper alternative to paying humans to moderate the same amount of content, it also allows the platform to grow. [[5]](#footnote-5)

**9.c. Government Regulations on Platform Policies**

While many platforms approached platform governance with self-regulation at the forefront, government legislation and rules can also form part of the governance associated with content available online. In 2021, the Australian federal Government introduced the *Online Safety Act 2021* (Cth) (*Online Safety Act*). Further regulations came into force with the *Online Safety (Basic Online Safety Expectation) Determination 2022* (Cth) *(Online Safety Expectations)*.

**9.c.i. *Online Safety Act -* Online Content Scheme**

The *Online Safety Act* is aimed at keeping consumers safe when accessing online materials or using platforms in addition to removing and reporting harmful content. The *Online Safety Act* established an Online Content Scheme. The Scheme is designed to ensure the aims of the Act are maintained and enforced through various measures.

The Online Content Scheme allows users to make formal complaints to eSafety about content that may be harmful, offensive and illegal. The Scheme also empowers eSafety to access and review the complaints.

The *Online Safety Act* part 9 has a classification system for content that may be illegal or restricted. Legal tools are available to keep Australian platform users safe, including:

* For class 1 material – involvement of the Australian Federal Police, and
* For class 2 material – an order for the removal or a restricted access system of the content.

The *Online Safety Act* can also empower eSafety to order the removal of extremely harmful and illegal content from platforms, even when it is not hosted within Australia. Extremely harmful and illegal content includes:

* Child sexual abuse material,
* Detailed instruction or promotion of crime or violence,
* Gratuitous, exploitative and offensive depictions of violence or sexual violence, and/or
* Material that advocates carrying out a terrorist act.

**9.c.ii. *Online Safety Expectations***

The *Online Safety Expectations* set out a range of expectations that online platforms are expected to conform to, aimed at increasing transparency and accountability of online service providers, and guaranteeing that platforms have adequate processes in place to minimise and mitigate harmful content.

1. ‘What is Governance’, *Governance Institute of Australia*, (Web Page, 1 September 2024) < https://www.governanceinstitute.com.au/resources/what-is-governance/>. [↑](#footnote-ref-1)
2. Suzor and Gillett, ‘Self-regulation and Discretion. Digital Platform Regulation’ (2022) *Palgrave Global Media Policy and Business,* 258-260. [↑](#footnote-ref-2)
3. Jonas Strandell, ‘What is Content Moderation? (Plus Best Practices)’, *Besedo* (Blog, 27 May 2024) < https://besedo.com/blog/what-is-content-moderation/> (“What is Content Moderation”). [↑](#footnote-ref-3)
4. Janselle M, ‘AI Content Moderation Vs. Human Content Moderation’, *Magellan Solutions* (Blog, 7 June 2024) < https://www.magellan-solutions.com/blog/ai-vs-human-content-moderation/> (“AI v Human Moderation”). [↑](#footnote-ref-4)
5. Ibid. [↑](#footnote-ref-5)