Public Consultation Feedback for

Draft Online Safety (Relevant Electronic Services – Class 1A and 1B Material) Industry Standard 2024
and
Draft Online Safety (Designated Internet Services – Class 1A and 1B Material) Industry Standard 2024

by
Tremau
www.tremau.com
Contact Person: Theodoros Evgeniou, [redacted]
General Tremau Inquiries: info@tremau.com

Attn: Executive Manager, Industry Regulation and Legal Services, Office of the eSafety Commissioner, Australia

This report provides **Tremau**'s (specifically, Theodoros Evgeniou – Professor at INSEAD and Co-Founder of Tremau) feedback to parts of the discussion paper *"about the establishment of two new industry standards under the Act: Online Safety (Relevant Electronic Services – Class 1A and 1B Material) Industry Standard 2024 (the Relevant Electronic Services Standard) and the Online Safety (Designated Internet Services Standard – Class 1A and 1B Material) Industry Standard 2024 (the Designated Internet Services Standard)."*

The focus is specifically on questions 15-20 which are related to generative AI.

Question 15 eSafety is seeking to place requirements on service providers that are best-placed to prevent the use of generative AI features to create and disseminate class 1A and class 1B material. Does the proposal achieve this?

A key aspect of the proposal is the ability to operationalize the classification of "*A service will meet the definition of a high impact generative AI designated internet service if it is reasonably foreseeable that the model may generate high impact material*" (a key point underlined).

Given it is currently (and, likely, for the foreseeable future) technically infeasible to ensure that no class 1A/1B material is generated and later disseminated using such service providers, it is appropriate to focus on due process related obligations. Significant guardrails, testing and red teaming, robustness, monitoring and other risk management requirements need to be in place, but also further requirements that these providers continuously scan and use upcoming innovations that can help the prevention of illegal content generation may need to be added. Service providers in this category may need to (a) put in place organizational capabilities (teams, tools, processes, governance) that do perform such technology innovation scanning and adoption, (b) report periodically progress they make in this direction.

Moreover, technical solutions that can support the identification of AI generated content (e.g., watermarking and variations) should complement due process requirements. Monitoring processes (leveraging AI tools as well as community/user reports) need to also be continuously strengthened and complemented with "notice & action" type transparency reports.

Finally, practices such as red teaming and continuous feedback and model improvement mechanisms (e.g. "RLHF" and the likes) may need to be added.

Question 16 Do the draft definitions for a high impact generative AI designated internet service and machine learning model platform service capture the right services? Are there types of providers that should not be included or should be excluded?

Overall, this is comprehensive, however particular consideration should be given to potentially new, for example distributed, channels for users to access models (and model weights) that may then be easily modified (e.g., guardrails removed) and used inappropriately. For example, model weights can be shared online "piecemeal" and via de-decentralized channels (e.g., similar to how movies have been distributed in the past).

It may be important to constraint any "piecemeal" distribution of AI model weights and ensure that any model used is complete ("one block"), with guardrails well documented, proper documentations (e.g., "model cards"), licensing and liability requirements, and with maximum feasible transparency.

Question 17 The high impact generative AI designated internet services category only captures models that meet a high impact threshold. It must be reasonably foreseeable that a service can be used to generate synthetic high impact material that would be classified as X18+ or RC. Is this threshold:
(a) appropriate for differentiating high impact and therefore high risk models?
(b) sufficiently clear to enable service providers to assess whether or not they meet the definition?
eSafety welcomes views on alternative thresholds which may be more suitable.

There is an interplay between "the ability of an AI model to generate high impact material" and the way the model is trained. For example, when training data include potentially illegal (or borderline illegal) content, it may be more likely for the AI model to also generate such content. While this is not guaranteed, perhaps some consideration of the training data used can be taken into account when a model is considered "high impact".

Moreover, the number and, perhaps, "type" of users may also matter. For example, if the users are restricted to business users, perhaps the risk is lower.

Question 18 In relation to high impact generative AI designated internet services, do the proposed obligations (in particular, the section 21 obligation to 'detect and remove' and the section 23 obligation to 'disrupt and deter' child sexual abuse material and pro-terror

material) provide appropriate safeguards? Are there specific challenges to deploying these measures in a generative AI context?

A key question to consider is the balance between technology vs. human centered solutions. For the former, the cost of deploying tools to comply with the obligations need to be considered. As new tools emerge, and older ones become increasingly more cost efficient and effective, it may prove important that a frequent (e.g., 2x a year) review process of the costs is put in place. Regarding the latter ("human labor"), perhaps minimal requirements on both governance and operational activities can be included, ensuring minimal costs.

Note: This recent report may provide more relevant details: https://crfm.stanford.edu/2023/12/01/ai-act-compromise.html

Question 19 In relation to machine learning model platform services, do the proposed obligations (in particular, the section 23 obligation to 'disrupt and deter') provide appropriate safeguards? Are there specific challenges to deploying these measures?

Similar to question 18, but perhaps only for a subset of the obligations and potential solutions used.

Question 20 In relation to relevant enterprise providers, do the proposed obligations (in particular, the section 23 obligation to 'disrupt and deter') provide appropriate safeguards? Are there specific challenges to deploying these measures?

Similar to question 18/19. Moreover, existing (non-AI related) regulations/obligations (for example regarding product safety) can also be leveraged when it comes to this group of stakeholders. See for example some points made in this report: https://computing.mit.edu/wp-content/uploads/2023/11/AIPolicyBrief.pdf (and in general in some reports from MIT: https://computing.mit.edu/ai-policy-briefs/)