stability.ai

**Response to Draft Online Safety (Designated Internet Services) Standard**
January 2024



## Introduction

Stability AI welcomes the opportunity to respond to the eSafety consultation into the Draft Online Safety (Designated Internet Services) Standard ("Draft Standard"). Stability AI is committed to realising the potential of AI while mitigating the risk of misuse, and we commend the leadership of the Australian Government in developing a targeted and timely response to emerging threats. The following submission outlines several promising features of the Draft Standard as it relates to generative AI models and systems, and makes a number of recommendations to help ensure the Draft Standard is feasible, effective, and accurately represents the complex AI supply chain.[1]

## Background

Stability AI is a global company working to amplify human intelligence by making foundational AI technology accessible to all. Today, we develop AI models across a range of modalities, including image, language, audio, and video. These models are software programs that can help a user to

[1] We are sensitive to concerns about the Draft Standard for user privacy and end-to-end encryption. Here, we focus our comments on the Draft Standard as applied to High Impact Generative AI Designated Internet Services, Machine Learning Model Platform Services, and Enterprise Designated Internet Services that provide AI models.

create, edit, or analyse complex content. With appropriate safeguards, we release these models openly, sharing our software code along with the billions of distinctive settings or "parameters" that define the model's performance. That means everyday developers and independent researchers can integrate or adapt our models to develop their own AI models, build their own AI tools, or start their own AI ventures, subject to our ethical use licences.[2]

To date, our models have been downloaded over 100 million times by developers, and nearly 300,000 developers and creators actively contribute to the Stability AI online community.[3] Our family of image models, Stable Diffusion, underpin up to 80 percent of all AI-generated imagery.[4] These models can take a text instruction or "prompt" from a user and help to create a new image. In addition, we develop a suite of language models that can interpret, summarise, or generate text. These include highly capable large language models, compact language models, specialised models for software development, and models for underrepresented languages, including Japanese and Spanish. Our audio model, Stable Audio, generates high-quality soundtracks and was recently listed on the *TIME* Best Inventions of 2023. Building on this experience, we have developed video models that demonstrate new breakthroughs in video generation.[5] Further, we support academic research into scientific applications of AI.

We are committed to the safe development of AI. Stability AI was pleased to contribute to the Department of Industry, Science and Resources consultation on supporting responsible AI, and consultations with Minister Rowland on generative AI governance. In addition, we are partners to the White House *Voluntary AI Commitments*, the British Government's *Joint Statement on Tackling Child Sexual Abuse in the Age of AI*, and the Singapore Government's Generative AI Evaluation Sandbox; we participated in the first large scale public evaluation of AI models at DEF CON, facilitated by the White House, as well as the UK AI Safety Summit at Bletchley Park; and we engage with authorities around the world on the future of AI oversight.

**Response**

Stability AI commends eSafety for drawing attention to some of the most serious, tangible, and immediate risks arising from the misuse of generative AI. Below we offer our response to the provisions of the Draft Standard that engage AI developers and AI deployers.

[2] See e.g. the Open Responsible AI License (OpenRAIL) for Stable Diffusion, prohibiting a range of unlawful or misleading uses, available here and the Stability AI Acceptable Use Policy, available here. We use the term "open" to refer to any models with publicly-available parameters.
[3] Figures from Hugging Face and Discord, November 2023.
[4] Everypixel, 'AI Image Statistics', August 2023, available here.
[5] See e.g. Stability AI, 'Improving Latent Diffusion Models', July 2023, available here; Stability AI, 'Stable LM-3B Technical Report', October 2023, available here; Stability AI, 'Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets', November 2023, available here.

1. **The Draft Standard acknowledges the distribution of responsibilities in the AI supply chain**

The Draft Standard recognises the complex relationship in the AI supply chain between upstream technology (e.g. models) and downstream systems (e.g. deployed applications). Models are just one component in an AI system such as a chatbot or image generator. By itself, the model is simply a piece of software accompanied by a large number of settings adjusted through training. Like other software, the model must be deployed in an application on a computing device in order to analyse or generate content. In that environment, different actors perform different functions, ranging from: training a base model (the raw "engine" that understands complex patterns and relationships within a textual, visual, musical, or scientific dataset); fine-tuning the model for a specific use-case (such as conversational interactions); distributing the model; hosting the model on a computing service; developing a user-facing application that interacts with the model; and promoting that application to users. Each actor may have limited visibility or control over downstream activity.

Conflating these actors, or imposing liability rules that assume vertical integration or formal relationships between these actors, would have a chilling effect on collaborative research and open innovation in AI. Instead, Stability AI encourages authorities to ensure that accountability is distributed across the supply chain. While there is no single gatekeeper and no silver bullets, every actor can play a role in helping to mitigate the risk of misuse.

To that end, Stability AI welcomes the clear distribution of responsibilities across the three designated internet services ("DIS") defined in the Draft Standard: High Impact Generative AI DIS (which we understand to mean deployed applications that enable an end-user to produce high impact materials); Machine Learning Model Platform Services (which we interpret to mean repositories that distribute or make available AI models for download); and Enterprise DIS (which we understand to include developers that provide models or services for enterprises to incorporate in their own products or services). This taxonomy recognises the distinct roles and responsibilities of application providers, repositories, developers, and account holders. These actors have different opportunities to mitigate potential risks. For example:

1. **Application providers.** Application providers can implement filters to interdict prompts, inputs, and outputs to prevent a user accessing unlawful content. Application providers may apply provenance features such as watermarking or metadata to help downstream platforms identify generated content.[6]

[6] For example, Stability AI is applying an imperceptible watermark to images generated on our API, in addition to Coalition for Content Provenance and Authenticity (C2PA) metadata that indicates the image was generated with a particular AI model.

2. **Repositories.** Repositories that distribute or make available AI components, such as inference code or model weights, can withdraw models that pose an unacceptable and unmitigated risk of misuse. Repositories can monitor trends in developer activity.

3. **Model developers.** Model developers can take steps to mitigate potential risks by adjusting the quality or composition of training data; optimising models through fine-tuning or reinforcement learning to amplify or suppress certain behaviours for particular tasks; and evaluating models to understand and disclose their known capabilities and limitations.[7]

Further, the Draft Standard distinguishes between different kinds of model providers, who range from everyday developers to sophisticated commercial labs. Stability AI supports that approach. Open sharing of foundational technology plays a vital role in helping to promote transparency and competition in AI, and authorities should avoid stifling that grassroots innovation by conflating developers.[8] Instead, the Draft Standard imposes targeted requirements on models that are provided to enterprise customers under a commercial agreement: ss 6, 23(4). That approach helps to ensure that the models most likely to be deployed at scale to Australian users are captured by the Draft Standard without imposing those obligations on models released for research, collaboration, or experimentation.

**Recommendation.** Stability AI supports the proposed taxonomy of designated services in the Draft Standard, as applied to AI, including the obligations for applications (High Impact Generative AI DIS), repositories (Machine Learning Model Platform Services), and commercial model developers (Enterprise DIS). However, Stability AI urges further consultation with all stakeholders to ensure that mitigation and reporting obligations can be implemented in practice.

2. **The Draft Standard emphasises feasibility, but may pose a significant burden for everyday developers**

The Draft Standard includes a number of measures that help to ensure obligations are proportionate and pragmatic. In particular, Stability AI welcomes the inclusion of "feasibility" criteria, such as those for detecting and removing content, which take into account the extent of the risk and the reasonable cost of mitigation: ss 7, 21(3)-(5), 22(3)-(5). Likewise, Stability AI supports the criteria for "appropriate action", which considers whether an alternative measure is

[7] For example, since taking over the exclusive development of the Stable Diffusion family of models, Stability AI has implemented a number of mitigations, such as filtering training data for unsafe content, which can help to prevent the model from generating unsafe content. In addition, we subject our models to internal and external evaluation as appropriate, and take steps to mitigate residual risks through fine-tuning and other techniques prior to release.

[8] See e.g. Stability AI, 'Statement to the AI Insight Forum', 2023 available here.

proportionate and reasonably expected to reduce the risk (ss 12, 25(2)). In addition, obligations such as mandatory development programs are subject to user-based thresholds (s 24). Together, these qualifications help to ensure the Draft Standard is risk-based, outcome-oriented, and adaptable:

1. **Risk-based.** Obligations should be proportionate to risk. In environments with a higher likelihood of harm, mitigation may be more stringent, without imposing the same mitigations on lower risk services. The feasibility and appropriate action criteria help to ensure the Draft Standard accounts for differences in the risk profile of different DIS.

2. **Outcome-oriented.** Different services may mitigate risks in different ways. In general, the Draft Standard focuses on specific outcomes rather than prescribing specific measures that may be ill-adapted to, or infeasible for, a particular internet service.[9]

3. **Adaptable.** In this way, the Draft Standard is better able to adapt to emerging or novel mitigations. For example, research is progressing rapidly in areas such as content provenance, deepfake detection, and classifier-based moderation. Focusing on outcomes rather than specific measures will help to ensure the Draft Standard accounts for future developments in technology and practices.

However, certain obligations are not subject to feasibility tests or size thresholds, and may pose a significant compliance burden for small or emerging businesses. For example, ss 23, 29, 34, or 41 requirements are not subject to ss 7 or 12 limitations. For these purposes, the Draft Standard treats all DIS identically. It would hold an independent app developer to the same standard of mitigation and assurance as a multinational technology company. Plainly, the safety and compliance resources available to Meta or Google are significantly greater than those available to everyday developers, and a "one size fits all" approach to these obligations could set back venture creation and service innovation in the Australian market.

**Recommendation.** We encourage eSafety to embed feasibility criteria throughout the Draft Standard to ensure that compliance obligations are appropriate and adapted to different services based on their size, activities, and resources. That would help to avoid a chilling effect on everyday researchers and independent developers who intend to serve the Australian market.

### 3. The Draft Standard may capture an unworkable number of services and content

The breadth of the Draft Standard poses a number of challenges. In particular, the Draft Standard regulates certain kinds of harmful content that may be impossible to accurately identify in

[9] See, e.g. s 23(3) n 1: "if a provider lacks such visibility or control of certain aspects [of its technology stack] such that it cannot deploy all mitigations, it can rely on other systems, processes and technologies which are available."

practice. For example, the Draft Standard requires High Impact Generative AI DIS to take steps to detect and remove known pro terror material that has been verified by a recognised organisation: ss 6, 22(2). However, the Draft Standard also requires DIS to deter, disrupt, notify, and respond to *unknown* pro terror material, defined as any material that "indirectly counsels, promotes, encourages or urges the doing of a terrorist act" or "indirectly provides instruction on the doing of a terrorist act": ss 6, 15(2), 17(2), 23(2). Likewise, High Impact Generative AI DIS must take steps to mitigate the risk of "material that depicts, expresses or otherwise deals with matters of drug misuse or addiction in such a way that the material offends against the standards of morality, decency and propriety generally accepted by reasonable adults": ss 6, 18, 25.

These are complex and highly subjective determinations that are difficult to implement at scale. AI technology can help to build more sophisticated moderation tools, and we are confident that language, image, and audio classifiers will play a vital role in detecting a range of harmful content. However, an overbroad broad definition of "pro terror material", "drug-related material", or "crime and violence material" could make it difficult for service providers to identify instances, implement reliable and robust mitigations, or measure their compliance state.

In addition, the Draft Standard sweeps into scope a vast number of services and tools that may incorporate AI only incidentally. For example, the definition of High Impact Generative AI DIS is broad, capturing a range of services that may use AI in only negligible or incidental ways to edit content. The Draft Standard appears to bring within scope nearly any hosted service – from word processing to image editing – that incorporates AI features if it is "reasonably foreseeable" that the service as a whole could be used to help generate synthetic high impact material: s 6.

**Recommendation.** We encourage eSafety to refine the scope of regulated content and services, ensuring that (i) content is defined objectively, (ii) content is capable of identification and moderation in practice, and (iii) High Impact Generative AI DIS is defined to exclude incidental or ancillary uses of AI within a wider service.

**Conclusion**

Stability AI thanks eSafety for the opportunity to comment on the Draft Standard, and welcomes the careful approach to regulatory architecture. We are pleased to support the work of the Commissioner, and to continue sharing our experiences as a model developer and application provider.