INDIAN GOVERNANCE AND POLICY PROJECT

IGAP PRIMER

CHILD SAFETY AND CHILD PROTECTION EXCEPTIONS UNDER THE DIGITAL PERSONAL DATA PROTECTION FRAMEWORK

Authored by
Dedipyaman Shukla and Jhanvi Anam

Introduction

Minors (persons under the age of 18 years) constitute a large section of users of the internet (estimated to be close to 1/3rd of total internet users), with indications that some of the screen-time of Indian urban youth being in excess of 3 hours daily.
A significant amount of this time is spent on social media platforms and online free-to-play gaming platforms. While these platforms are frequently accessed by children, they have been associated with dangerous risks to physical and mental health through exposure to harassment, cyberbullying, sexual exploitation, exposure to content promoting eating disorders, privacy concerns, etc, among other harms. To remedy many of these real risks, child-centric regulatory frameworks have evolved alongside a range of technological solutions.

Regulatory Frameworks

Specific to data privacy, the Digital Personal Data Protection Act, 2023 *(DPDP Act)* and the Draft Digital Personal Data Protection Rules, 2025 *(Draft Rules)* mitigate the risks faced by children through legal interventions. The DPDP Act requires that the consent of a parent must be taken before the personal data of a child is processed by entities known as 'Data Fiduciaries' under the DPDP Act.* This requirement ensures that informed consent can be provided by an appropriate stakeholder. However, the DPDP Act also goes on to prescribe specific restrictions on how personal data of children can be processed further once parental consent is obtained.

* Under Section 2(i) of the DPDP Act, a data fiduciary means any person who alone or in conjunction with other persons determines the purpose and means of processing of personal data.

Restrictions on Processing Children's Data

IDENTIFICATION OF PARENT OF MINOR

SHARING OF NOTICE & OBTAINING CONSENT

DATA PROCESSING

SUBJECT TO FOLLOWING RESTRICTIONS

- No tracking
- No targeted advertising
- No behavioral monitoring
- Processing cannot cause a 'detrimental effect'

Under Section 9(2), the DPDP Act prohibits Data Fiduciaries from undertaking processing that is likely to cause a detrimental effect on the well-being of a child. Further, Section 9(3) prohibits tracking, behavioural monitoring, or targeted advertisement directed to children, subject to such specific conditions. To provide additional guidance on the extent of these requirements, certain exceptions have been provided under Rule 11 and the 4th Schedule of the Draft Rules, which permit the processing of children's data for specified purposes. Put simply, these exceptions may enable behavioral monitoring of children to prevent access to information that could be detrimental to a child's well-being. Such an exception may prove to be vital to the protection of children from nefarious digital activities, and provision a higher standard of online safety. However, the extent of these legislative exceptions may not be aligned with the variety of technological solutions that prevent online harms across different digital platforms.

In this context, we will discuss the role of various technologies already in use to protect children online, especially on social media and gaming platforms, in the context of the DPDP Act and Draft Rules.

Existing Protection Mechanisms

To mitigate the multitude of risks faced by children on digital platforms, a multitude of automated technologies have been deployed, which may rely on personal data processing.

Among these, some of the most prominent technologies/solutions are analysed below:

SR. NO.	TECHNOLOGY TYPE	FEATURES	DATA INPUT	OUTCOME OF USE	IMPACT BY DPDP
01	Text Classification approach through Machine Learning methods	Uses text classification to detect predatory behaviour in chat logs using a wide variety of well-known classifiers	Text, behavioural and demographic features in chat logs	Detection of predatory behaviour towards children and grooming	Text classification is conducted to detect predatory behaviour. However, it is unclear if this system is exclusively used to ensure that such predatory messages are not accessible to the child.
02	Sentiment Analysis	Attempts to quantify emotional labels, positive or negative attitudes (polarity classification) and emotional intensity from text	Chat logs	Filtering out 'trolling' messages and spam	This method helps to detect cyberbullying messages in text. If incorporated within gaming systems, it filters such messages and prevents children from seeing such messages. Therefore, it may fall within the scope of the Draft Rule exemption.
03	Hybrid Deep Learning approach (DEA-RNN)	By leveraging 'neural networks' and optimization algorithms, these systems are able to adapt to the dynamic nature of social media platforms and improve the accuracy of cyber-bullying detection over time	Short-text formats (eg: tweets)	Detection of cyberbullying in the Twitter dataset	While DEA-RNN was experimentally tested only on Twitter datasets, it is unclear if this system can be exclusively used to ensure that such predatory messages are not accessible to the child.
04	AI-Driven Voice Chat Detection (Safe Voice by Unity)	Leverages cutting-edge acoustic and semantic intelligence in order to identify in-game toxicity	Digital sound-recording / acoustic signature	Provides detailed session reporting dashboards to moderators on toxic behaviour	While this method helps identify toxic in-game behaviour, moderators are empowered to escalate incidents with discretion.
05	Trust Factor System (By Valve)	Applies a factor to competitive in-game matchmaking based on likelihood of a player to receive a ban	Player metrics associated with their Steam account for gaming	Improved matchmaking among players perceived as 'good' actors	While such a system may have potential application in the context of preventing exposure to harmful content, it is deployed for a wide variety of purposes, including the promotion of 'non-toxic user' behaviour and discouraging cheating.

Note: The text classification approach was primarily developed through available data sets to detect predatory behaviour in chat logs by using a variety of classifiers which utilise traditional ML methods. This is the most widely used technique for detecting predatory behavior on chatting platforms. Sentiment Analysis and Hybrid DEA-RNN model are examples of classifiers. Sentiment Analysis is used to determine the attitude or emotional reaction of a writer, one such example is that of Microsoft's Azure sentiment analysis. Hybrid deep learning approach, DEA-RNN is a proposed model that works to detect cyberbullying utilizing textual content of tweets, whereas the other type of media such as images, video, and audio is still an open research area. The aim of the proposed model is to classify and detect cyber bullying tweets in a real-time stream.

Summary

While existing technological solutions effectively mitigate many online harms faced by children, the alignment of these tools with the DPDP Act and the Draft Rules remains ambiguous. Further, the exceptions provided under the 4th Schedule of the Draft Rules allow for behavioural monitoring to prevent access to harmful content but lack adequate clarity.

> *The language of the exception under the 4th Schedule reads: 'For ensuring that information likely to cause any detrimental effect on the wellbeing of a child is not accessible to her'*

Neither the DPDP Act nor the Draft Rules provide clear illustrations which demonstrate the elements of "detrimental effect".

RECOMMENDATIONS

- Refining the Exception Clause in the Draft Rules – The language of the exception in the 4th Schedule should be refined to explicitly allow the deployment of child protection technologies while ensuring compliance with privacy safeguards. This may be framed as a flexibility to the Data Fiduciary to provide customizations in the best interest of the child, ensuring that protective mechanisms like behavioural monitoring and filtering harmful content can be implemented responsibly.

- Definition of "Detrimental Effect" – The Draft Rules should provide a clear and illustrative definition of what constitutes a detrimental effect on children's well-being to remove regulatory uncertainty.

INDIAN GOVERNANCE AND POLICY PROJECT

The Indian Governance And Policy Project (IGAP) is an emerging think tank focused on driving growth, innovation, and development in India's digital landscape. Specializing in areas like AI, Data Protection, FinTech, and Sustainability, IGAP promotes evidence-based policymaking through interdisciplinary research. By working closely with industry bodies in the digital sector, IGAP provides valuable insights and supports informed decision-making. Core work streams include policy monitoring, knowledge dissemination, capacity development, dialogue and collaboration.

For more details visit: www.igap.in

relations@igap.in | igap.in