

VENDREDI 3 OCTOBRE 2025

HACK THE DSA

*Les rapports sur les risques
systémiques*



chaire 
sur **la modération**
des *contenus*

Problèmes de base

- > Comparaison des rapports
- > Manquements de ces rapports
- > Voies d'amélioration



RAPPORTS RISQUES
SYSTEMIQUES



PROBLEMES DE
FOND

PROBLEMES DE
FORME

Constats et observations - Forme

NEGATIF

1. Dense et peu digeste
2. Absence de standardisation dans la méthode d'analyse et la présentation
3. Caviardage

POSITIF

1. Amélioration de la mise en page
2. Présence de matrices
3. Mention des articles DSA



	• Article 35(1)(c): X will continue to iterate and improve upon automated moderation techniques for improved detection of violative content before it is reported. [REDACTED]
Risk of dissemination of illegal content	[REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED]
Risks of negative effects to fundamental rights	[REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED]



IX. Annex: Matrices

1. Probability matrix

	1	2	3	4	5
Probability	Very unlikely	Unlikely	Possible	Likely	Almost certain
Frequency of incident or event occurring	May occur within a year Rare but could occur	May occur within 6 months Has occurred for comparable platforms	May occur within a month Has occurred for X and / or commonly occurs for comparable platforms	Likely to occur within the next 2 weeks Has occurred for X regularly	Immediately or within days Occurs for X every day

Fig.13: Probability scale for the purpose of the DSA risk assessment

2. Severity matrix

	1	2	3	4	5
Severity	Very low severity	Low severity	Moderate severity	High severity	Very high severity
Scope of impact: The extent to which the harm is physical, psychological, informational, economic, and/or societal; as well as how the harm may be experienced by vulnerable groups. weighed at 50%	Very low harm on the populations impacted by the risk.	Low gravity of harm on the populations impacted by the risk, such as informational, economic, and/or societal harm.	Moderate gravity of any harm, especially physical and/or psychological harm, on the population impacted by the risk. Harm specifically impacts vulnerable groups	High gravity of any harm, especially physical and/or psychological harm, on the population impacted by the risk. Harm specifically impacts vulnerable groups	Very high gravity of any harm, especially physical and/or psychological harm, on the population impacted by the risk. Harm specifically impacts vulnerable groups
Scale/reach: Number of users affected in the EU weighed at 40%	Reaches a negligible number of users	Reaches a minimal/minor number of users	Reaches a moderate number of users	Reaches a high number of users	Reaches most users of the platform
Remediability: Reversibility of the harm or difficulty in restoring the situation weighed at 10%	Remediable: Remedy will restore the person/situation to the state before the impact.	Likely remediable: Remedy is likely to restore the person/situation to the state before the impact.	Possibly remediable: Remedy may help to restore the person/situation to the state before the impact.	Rarely remediable: Remedy can rarely restore the person/situation to the state before the impact.	Not remediable: Remedy cannot restore the person/situation to the state before the impact.

Fig.14: Severity scale for the purpose of the DSA risk assessment

3. Inherent risk matrix

	Severity					
		Very low severity	Low severity	Moderate severity	High severity	Very high severity
Probabil ity	Almost certain	Low	Medium	High	Critical	Critical
	Likely	Negligible	Low	Medium	High	Critical
	Possible	Negligible	Low	Low	Medium	High
	Unlikely	Negligible	Negligible	Low	Low	High
	Very unlikely	Negligible	Negligible	Negligible	Negligible	High
	Adjustment has been made for very high severity events, such as crisis events, where the risk is higher than usual. Thus, for this situation, the risk has been increased to high risk.					

Fig.15: Residual risk matrix for the purpose of the DSA risk assessment

Constats et observations - Fond

1. Autopromotion / Plaidoyer

- Absence d'autocritique sur les risques internes au service
- Valorisation des mécanismes d'atténuation sans tierce vérification possible (données non disponibles)

2. Données

- Absence
- Superficialité
- Subjectivité



Constats et observations - Fond

3. Présentation trompeuse

- Présentation manipulatoire (saturation d'information, cherry picking)



Risk Category	Risk Scenario	Controls & Measures in place	Additional (mitigation) Measures to implement
Protection of Minors	Inadequate access verification systems allowing minors access to adult content	- warnings about adult content - age confirmation - guidance for guardians for protection of minors - security measures and parental controls setup - Terms of Service - Point 2 - Access (https://info.xvideos.net/legal/tos/)	- consider liaising with UNICEF and other globally operating NGOs with similar impacts to fund programs focused on raising awareness of on-line safety and security of minors - consider expanding and improving the wording of the disclaimer about the inappropriateness of content for minors on entry and page - closely monitor the latest developments in age verification tools with a view to finding an instrument that strikes a sufficient balance between effective protection of minors and the fundamental right to do business - support the development and raising awareness of effective age verification tools, such as those integrated into the operating system of individual devices - consider liaising with experts and NGOs dealing with the protection of minors and age verification tools

There are more than 100,000 app submissions in an average week. In 2022, App Review reviewed 6,101,913 submissions (including app updates), of which over 25 % were rejected by the App Review team for various compliance issues.²⁰ In 2023, App Review reviewed 6,892,500 submissions (including app updates); again, over 25 % were rejected.²¹ App

While TikTok does not believe that use of platforms (in general or of TikTok in particular), is automatically harmful in and of itself, it recognises that some users (in particular Younger Users) could at times and in specific circumstances face serious negative mental and physical health impacts as a result of the way in which they use such platforms.

Given the volume of content on the Pinterest platform, detecting and taking appropriate action on harmful content cannot completely eliminate these four risk areas. There are inherent risks on any platform that deals with vast quantities of content and data. Pinterest takes a risk-based approach when it comes to content moderation, prioritising harms that pose the greatest potential risk and balancing fundamental rights with keeping users safe.

Recommandations

1. Standardisation des méthodes
2. Ajout de données sourcées et vérifiables
3. Contrôle de la Commission Européenne (lignes rouges)
4. Partage de bonnes pratiques inter-acteurs (société civile – services numériques – régulateurs)

