Brief for the Hearing at the European Parliament on DSA Enforcement and the Protection of Minors

Significant momentum is building around Europe to step up on online minors protection.

In October last year the European Parliament adopted a strong own-initiative report on that matter, urging the European Commission to make "full use of its powers under the Digital Services Act (DSA), including issuing fines or, as a last resort, banning non-compliant sites or applications that endanger minors". MEPs also proposed an EU-wide digital minimum age of 16 for access to social media, video sharing platforms and AI companions, unless authorised by parents.

Two weeks ago the European Commission preliminarily found TikTok in breach of the DSA for its addictive design. According to these findings, TikTok did not adequately assess how its addictive features could harm the physical and mental wellbeing of its users, including minors and vulnerable adults. It is the first case in which the Commission used articles 34 and 35 of the DSA to question fundamental design choices made by the company.

And this is exactly what the European public expects from EU institutions today. Recent YouGov polling showed 70% of Europeans support the Commission escalating measures against X if it fails to address breaches to EU law. Almost half want a full ban of the platform in Europe.

Are very large online platforms (VLOPs) visibly concerned about this shift? In their Risk Assessment reports (last round was published in 2025) companies running the largest social media platforms – TikTok, Meta and YouTube (Google) – declared a long list of mitigation measures meant to make social media safer for kids. But, once again, none of them disclosed quantitative or qualitative data to substantiate the effectiveness of their mitigation measures. Which is why we can't really tell how many of these measures bring expected results and whether they were designed to bring the results we would wish for. But we have been able to find evidence that suggests otherwise[1].

In this brief, prepared for the purpose of the hearing organised by EPP and Renew on DSA Enforcement and the protection of minors, we tried to match the most promising claims made by VLOPs with evidence coming from independent research or civil society investigations.

[1] Evidence quoted in this brief comes from CSOs and researchers who are part of DSA Civil Society Coordination Group, led by CDT, and Recommender Systems Task Force, led by Panoptykon. Each piece of work has been attributed to its authors in the main text.

We chose to focus on TikTok, Meta and YouTube – as owners of the most popular social media platforms. We hope that this brief will be useful for the European Parliament and European Commission in their work related to DSA enforcement. Due to lack of access to relevant metrics owned by companies themselves, at this point we can't say that mitigation measures declared by VLOPs in their reports are not effective. But we are convinced that, in the light of evidence casting doubt on the effectiveness of some measures, the burden of proof is now on VLOPs to prove us wrong. In all honesty, we wish that data revealed by the companies prove us wrong one day, as it would be great news for their users and a milestone for VLOPs' public accountability. With this intention, we invite you to check our selection of – what seemed at first sight – very promising mitigation measures, and the questions that should be answered by their designers.

1st theme:
Mitigating risk of excessive platform use, prompted by dark patterns or addictive design

In their reports, very large social media platforms acknowledge the risk that their interface, design, or features may stimulate behavioural addictions in minors using the service, and reassure us that their designers "are working on it". How? Mainly through sleep reminders, screentime break prompts (switched on by default for younger users), take-a-break videos and campaigns promoting "healthy online habits".

One may wonder why, at the same time, all VLOPs **continue to use hyper-personalized feed, combined with infinite scroll, auto-play and virality signals like the 'like button'**, in spite of mounting scientific evidence that problematic use of social media leads to sleep and attention problems and even changes in brain structure similar to those observed in people experiencing drug addiction. Is it because the company's topline metrics (set by the management) still put short term "user engagement" above safer, healthier and – in the longer run – better user experience, even for the youngest users?

The table below shows what VLOPs tell us in their Risk Assessment reports about mitigating risk of excessive/problematic platform use (1st column), what we – CSOs – see through the lens of independent research (2nd column), and what we would really like to hear from VLOPs (what statistics and metrics we should like to see) in order to get a full picture (3rd column):

Quote from VLOPs' Risk Assessment reports (2025)	Evidence and/or expert arguments challenging the effectiveness of the mitigation strategy
TikTok "TikTok implements a number of mitigation measures to address Online Engagement Risks, including a daily screentime limit for Younger Users that is set to 60 minutes by default, Family Pairing controls which allow parents/guardians to manage the time they spend on the Platform, sleep reminders, screentime break prompts, and take-a-break videos."	Amnesty International's research on TikTok shows that the company has maximized the addictive qualities of design choices and engagement strategies employed by competing social media companies, incentivising a race to the bottom between a small number of leading social media companies vying for the highest user numbers and engagement rates. TikTok has done this in spite of mounting scientific evidence of the serious risks associated with addictive use of social media especially for children and young people's health, including sleep and attention problems and even changes in brain structure similar to those observed in people experiencing drug addiction. Testimonies from children and young people were corroborated by specialist adolescent psychologists consulted as part of Amnesty International's 2023 and 2025 research, who said that they had observed a trend of TikTok aggravating existing issues with addictive patterns in children's and young adults' use of social media. Neurologist Servane Mouton further explained the mechanisms and effects in a 2025 interview with Amnesty International: "The economic model of social networks is based on stimulating the short-term reward system so that we go there as often as possible, for as long as possible. It is a bit like chasing after dopamine shots that will immediately produce the sensation of pleasure and make us want to go back" **Amnesty International, Driven Into Darkness, 2023** **Amnesty International, Dragged into the Rabbit Hole, 2025** Documents released in US litigation show that TikTok leadership approved the rollout of screentime interventions only on the condition that the tools would not have significant impacts for heavy users. US plaintiffs cite internal communications from TikTok's global Research and Development Chief, suggesting that TikTok can "accept a 5% drop in stay time for Screen Time

	Management features for special user groups like minors and excessive users." **Knight Georgetown Institute (KGI), Measuring Risk: What EU Risk Assessments and US Litigation Reveal About Meta and TikTok, 2026**
YouTube "We also assessed the risk that the interface, design, or features of YouTube stimulate behavioural addictions in minors using the service. (...) Google has many measures in place to address this risk such as: • parental controls • the unique experiences designed for kids • and surfacing high-quality content." "We have disabled a number of standard features normally available in YouTube, like comments, uploads, purchases, and live chat. To reinforce healthy screen time habits, reminders for breaks and bedtime are set to "on" by default."	A survey of 1,500 young people in Europe found that 68% of children said they felt 'stuck' on YouTube at least once a week (including 40% everyday). 59% of those children say they lost sleep weekly because of this. When kids were asked why they felt stuck on YouTube, **84% said it was the endless feed, and 76% said it was the personalisation of content.** **Reset Tech, 2025 study** "Providers and CSOs have identified mental and physical health risks associated with the design and use of VLOPs and VLOSEs, particularly in relation to excessive, compulsive and addiction-like use of social media. Some **noted the role of platform design features, such as infinite scroll, autoplay, ephemeral content notifications** and other interface elements linked to patterns of compulsive use" **European Board for Digital Services, 1st annual report, 2025**
Meta "Meta has a taxonomy of design patterns that have been identified as likely to be deceptive in order to avoid implementing such designs on Meta's platforms. The taxonomy utilises patterns flagged as noteworthy by key external sources (e.g., DSA…) to provide examples that represent deceptive designs that may lead a reasonable person to feel tricked, misled, coerced, or unduly pushed into doing something they wouldn't have otherwise chosen to do. The taxonomy is used by product designers	"1.5 million US teen users, or 11.7% of teen users, exhibit one problematic use behavioral proxy, defined as: 1. No-engagement late night session; 2. Late night high-volume product switching; 3. No-engagement repeat sessions (Less than 10 min from end of prior session); 4. High volume of short sessions (< 15 sec); 5. High Frequency notifications checking. When combined with overall nighttime use, the internal presentation estimates "that 18.3% of IG [Instagram] teen users exhibit behavior associated with problematic use." **KGI, Measuring Risk…, 2026** During our testing we did not note any meaningful changes to design elements aimed at increasing engagement and time spent. Although there are now screen time reminders, our avatars

as guidance for their designs to avoid deceptive designs." "Meta takes steps to embed safety by design to help users engage safely online and on our products, particularly for minors."	remained subject to **persistent notifications, autoplay, infinite scrolling, and prompts to interact with suggested content.** Even after logging off, our accounts received app notifications. **5Rights, 2025 study** **Platform design plays an important role in creating risks of unwanted and harmful contact.** Research has found that expansive default account visibility and account recommendations are crucial design vulnerabilities for sextortion targeting minors. **KGI/Panoptykon, submission to EBDS, 2024** "(...) companies like Apple and Google (...) **could empower people to set predictable times during the day or week for when they want to check "slot machine" apps**, and correspondingly adjust when new messages are delivered to align with those times." "News feeds are purposely designed to **auto-refill with reasons to keep you scrolling**, and purposely eliminate any reason for you to pause, reconsider or leave. It's also why video and social media sites like YouTube or Facebook **autoplay** the next video after a countdown instead of waiting for you to make a conscious choice (in case you won't)." "Facebook Messenger (or WhatsApp) (...) **design their messaging system to interrupt recipients immediately** (and show a chat box) instead of helping users respect each other's attention. By default Facebook **tells the sender when you saw their message**, instead of letting you avoid disclosing whether you read it." **Tristan Harris, How Technology is Hijacking Your Mind, 2016**

Relevant metrics we expect all VLOPs to share (by Integrity Institute) (*)

Scale

- How does the platform define 'objective harmful use' and 'unbalanced engagement'? How many EU teens exhibit problematic use? (however defined)
- What fraction of teens exhibit problematic use?
- How many instances of problematic use are there?
- What is the distribution of hours spent per week for teen users?
- What is the distribution of hours spent per week during sensitive time periods (i.e. during school hours, during sleeping hours)?
- Summarize the results of relevant A/B tests related to overall time spent on the platform following activation of screentime management tools (by the user and/or the parent).

Cause

- How many teens use features meant to control problematic use? (parental controls, screen limits, etc.)
- What % of all teens on the platform use these features?

- What is the distribution of hours spent per week for teens that use these features? How does it compare to teens who don't use any features to control problematic use?
- What is the distribution of hours spent per week during sensitive time periods (i.e. school hours, sleeping hours) for teens that use these features? How does it compare to teens who don't use any features to control problematic use?

Nature

- How does the distribution of time spent per week differ for teens and adults? How does the distribution of time spent per week during sensitive time periods differ for teens and adults?

Design

- What fraction of time spent for teens goes to various portions of the app? (discovery feed, friends feed, messaging, searching, etc.)
- What fraction of time spent during sensitive times (school, sleep) for teens goes to various portions of the app? (discovery feed, friends feed, messaging, searching, etc.)
- What is the breakdown of entry points to the platform for sessions that involve use during sensitive time periods? (What % is from notification, external link, person opening the app, etc)
- What are all the ways in which the platform systems predict how much time users will spend with the app?
- What systems and features are designed to maximize the time spent? Please show results of testing your design choices that affect time spent (both positively and negatively).
- [for TikTok] What fraction of users watch the full take a break video? What percentage watch half or less? What percentage watch 5 seconds or less?

Governance

- How did they come to its definition of "problematic use"? (What child psychology experts or CSOs etc. did they talk to? What survey/focus group etc research did they perform on teens to assess?
- Are there topline metrics the company uses and goals that incorporate time spent? Are there goals around increasing time spent on the platform?
- How often do efforts to control problematic use impact those metrics and goals?
- How does the company manage when there are tensions between efforts to control problematic use and goals around their topline metrics?
- [for TikTok] Is it still true that TikTok teams are not allowed to reduce overall time spent on the platform by more than 5%? [source: KGI, Measuring Risk: What EU Risk Assessments and US Litigation Reveal About Meta and TikTok, p.12-13]

(*) stats and numbers should be broken down by country

2st theme:
Mitigating the risk of minors' exposure to harmful content by re-designing algorithmic systems

In their Risk Assessments all VLOPs advertise product features that are meant to prevent minors from being exposed to illegal, harmful or age-inappropriate content. Mitigation measures include: signals for estimating the age of users, age-restricted content, and – the most valuable promise--avoiding recommendations that could be low-quality, objectionable, or particularly sensitive.

YouTube (Google) goes as far as to declare that – instead – their content curation algorithms use meaning, relevance, quality and usefulness to prioritise the results that seem most helpful.

Yet, independent research shows there are minimal differences in content exposure between adult and youth accounts, raising concerns about the effectiveness of age-based content curation. Experiments run with avatars of teenage users show that they continue to receive recommended posts focusing on subjects they had never indicated an interest in. And, most disappointingly, underage users continue to report rabbit-holes full of depressive, suicidal and other toxic content.

The table below shows what VLOPs tell us in their Risk Assessment reports about mitigating risk of excessive/problematic platform use (1st column), what we – CSOs – see through the lens of independent research (2nd column), and what we would really like to hear from VLOPs (what statistics and metrics we should like to see) in order to get a full picture (3rd column):

Quote from VLOPs' Risk Assessment reports (2025)	**Evidence and/or expert arguments challenging the effectiveness of the mitigation strategy**
YouTube "(...) the service type and corresponding risk profile remain key factors in determining the greatest inherent risks, highly motivated bad actors are a primary concern, and we continue to invest in programs aimed at reducing levels of residual risk."	**Methods used by VLOPs to control age are not effective:** "Recommender systems were noted (...) for their potential to amplify the dissemination of harmful content (e.g. legal but potentially problematic or content being algorithmically recommended to child users with increasing frequency and thus becoming harmful for its cumulative effect)." **European Board for Digital Services, 1st annual report** "Despite evidence that the core mechanics of personalized recommendations, virality and "optimizing" to maximize user engagement not only create risks of human rights abuses and violations, but have contributed to concrete harms, social media platforms such as TikTok, Instagram and Facebook continue to employ them to fulfil their goal. Indeed, TikTok's success in doing this has pushed its competitor Meta (Instagram and Facebook's parent company) to emulate TikTok's reliance on personalized recommendations within its engagement strategy and to invest further in refining its algorithmic systems." **Amnesty International, Submission to the DFA call for evidence by the European Commission, 2025**
TikTok "TikTok prohibits a range of edited or AIGC content that is misleading or that contains certain depictions of minors, which mitigates the risk of […redacted…]. Restrictions on the use	25% of TikTok's top search results contain synthetic AI imagery vs. significantly less on Instagram. Over 80% of AI content comes from Agentic AI Accounts on TikTok. Only half of TikTok's AI content receives proper labeling; 23% on Instagram.

of AIGC also lower the risks of potentially negative effects from a proliferation of content depicting unrealistic beauty standards."	Over 80% of AI content is photorealistic, increasing deceptive potential. **AI Forensics,** 2025 study
TikTok "TikTok's *Mental and Behavioural Health* policy prohibits or age-restricts a range of content related to suicide and self-harm, disordered eating and body image and dangerous activities and challenges."	39% of 7th grade students in Poland (age) states that they are receiving **suicide-related content** on social media while not seeking that content. 57% of 7th grade students in Poland (age) who had previously attempted suicide are receiving suicide-related content on social media while not seeking that content. **BGK,** 2025 study "We routinely detect pro self-harm and pro-anorexia content that is accessible to minors as part of our content moderation experiments. The only available write up of this so far is from Australia, where we found 115 pieces of content (and found TikTok moderated only 15% when reported)." **Reset Tech,** 2025 study
TikTok "TikTok uses content classification measures (Content Levels) to organize content and prevent Age-Restricted Content (i.e. content that does not violate the Community Guidelines, per se, but that TikTok considers is only appropriate for Adult Users due to its complexity and overt maturity) from reaching Younger Users."	TikTok's search suggestions were highly sexualised for users who reported being 13 years old and had "Restricted Mode" turned on. TikTok suggests around 10 searches that "you may like". The search suggestions provided to our test accounts often implied sexual content. For three of our test accounts, sexualised searches were suggested the very first time that the user clicked into the search bar. For all seven users, we encountered **pornographic content** just a small number of clicks after setting up the account. This ranged from content showing women flashing to hardcore porn showing penetrative sex. The **platform wasn't just showing such content to a minor, but actively directing them to it** when the account user had zero previous search or watch history. **Global Witness,** 2025 study
YouTube "We age-restrict content that does not violate our policies, but is nonetheless inappropriate for viewers under 18. This includes videos containing adults participating in dangerous activities that minors may imitate or videos related to regulated substances,	"We collected over 7000 videos, classified them as harmful vs not-harmful, and then simulated interactions using age-specific sockpuppet accounts through both passive and active engagement strategies. We also evaluated the performance of large language (LLMs) and vision-language models (VLMs) in detecting harmful content, identifying key challenges in precision and scalability. Preliminary results show **minimal differences in content exposure between adult and youth accounts**, raising concerns about the platform's age-based moderation. These findings suggest that the platform needs

sexually suggestive content, or violent and vulgar content. Videos that are age-restricted are not viewable by signed-out users either."	to strengthen youth safety measures and improve transparency in content moderation." **Towards an Automated Framework to Audit Youth Safety on TikTok, Linda Xue, Francesco Corso, Nicolo' Fontana, Geng Liu, Stefano Ceri, Francesco Pierri, 2025** "A comparison between 13-year-old and 18-year-old user accounts shows that minors face disproportionately higher levels of harmful videos, spanning 7–15% versus 4–8% for adults. On YouTube, 15% of recommended videos to 13-year-old accounts during passive scrolling were assessed as harmful, compared to 8.17% for 18-year-old accounts. On TikTok: it was 7.83 % of recommended videos to 13-year-old accounts vs to 5.67 % to 18-year-old accounts." "Overall, children commonly encounter harmful content in under five minutes, compared to roughly nine minutes for adults, raising concerns about the current safeguards' capacity to prevent early, potentially harmful exposure." **Protecting Young Users on Social Media: Evaluating the Effectiveness of Content Moderation and Legal Safeguards on Video Sharing Platforms, F. Eltaher et al., 2025 (preprint)**
Meta "Through the Recommendations Guidelines, Meta works to avoid making recommendations that could be low-quality, objectionable, or particularly sensitive, and/or also inappropriate for younger viewers."	"Instagram immediately suggested our avatar accounts to connect with adult-owned accounts. While some restrictions exist – such as preventing adults from messaging minors unless they follow each other – our testing found that teen avatars could also still freely send message requests to adults. The algorithm routinely encouraged our avatar accounts to widen their network of contacts with unknown strangers, increasing potential privacy and safety risks." **5 Rights, 2025 study**
Meta "Meta implements age-related recommendation restrictions to reduce the likelihood of minors encountering potentially sensitive or low quality content."	Instagram's 'Discover' function recommended harmful and misleading content to our teen avatars. This included fad weight loss 'challenges' and posts containing medical misinformation. Instagram immediately suggested our avatar accounts to connect with adult-owned accounts. Avatars were recommended posts focusing on subjects we had never indicated an interest in. **5 Rights, 2025 study**

TikTok "TikTok uses dispersion techniques to reduce the risks associated with Concentrated Content, by dispersing content containing certain themes, including borderline extreme dieting and fitness, borderline negative affect, adult nudity and sexual activity and mental health narratives, which may not be problematic when viewed occasionally, but may become so if viewed repeatedly."	TikTok's recommender system can draw teen accounts into a rabbit-hole of depressive content in less than an hour of use once they signalled an interest in 'sadness'. **Amnesty International, Driven Into Darkness, 2023** TikTok's 'For You' feed poses additional risks for children and young people with pre-existing mental health concerns. A technical investigation conducted in 2023 by Amnesty International, the Algorithmic Transparency Institute (National Conference on Citizenship) and AI Forensics shows that children and young people who watch mental health-related conten on TikTok's 'For You' page can easily be drawn into "rabbit holes" of potentially harmful content, including videos that romanticize and encourage depressive thinking, self-harm and suicide. In 2025, Amnesty International repeated its research (with technical support from the Algorithmic Transparency Institute (National Conference on Citizenship)) into rabbit holes of harmful depressive, self-harm and suicide content focusing on French-language content available through the app in France. Renewed qualitative and quantitative evidence reveals TikTok's continued failure to address the risks and harms caused and contributed to by its business model. Testimonies reveal how drawing them into an artificial tunnel vision fixated on mental health struggles, TikTok normalized and exacerbated their self-harm and suicidal ideation up to the point of recommending content on suicide methods and challenges. **Amnesty International, Submission to the DFA call for evidence by the European Commission, 2025** **Amnesty International, Dragged into the Rabbit Hole, 2025**
YouTube "We keep users and society safe through built-in protections (...) to prevent, detect, and respond to illegal and harmful content. In Q1 2025, we removed 54.67% of violative videos before they had a single view, and 27.28% of violative videos when they had one to ten views.	YouTube does not consistently remove illegal hate speech, even when they are made aware of it. In a 6 month experiment, content classified as illegal hate speech by lawyers was only removed 66.2% of the time after user-reporting. **Reset Tech, 2025 study** YouTube has the lowest removal rate of illegal content of all providers examined. Overall, only 32% of reported content was removed during the period under review. While 38% of content was removed in 2024, this figure fell to 28% in 2025. It is striking that YouTube did not communicate any decision on a significant part of reports even though the platform acknowledged the receipt of all reports: in 2024, this affected almost half (48%) of cases, and in 2025, around a third (34%). **HateAid, 2025 report**

Our VVR reports indicate that violative views today are around 0.1% of all videos viewed (i.e., out of every 1,000 views on YouTube, just one is of violative content)."	
YouTube "Our systems use meaning, relevance, quality, usability, context and hundreds of other signals to prioritise the results that seem most helpful, in particular content that seems to demonstrate expertise, experience, authoritativeness, and trustworthiness." "Over the past several years we have invested significantly in the recommendations systems where accuracy and quality are key, including news, politics, and medical information. Our systems are trained to elevate high-quality sources in search results, particularly in sensitive contexts, and we provide high-quality information in information panels, in turn helping people find accurate and useful information."	76% of video views on YouTube are driven by recommendations. YouTube amplifies negative emotions, by increasing their prevalence and prominence in recommendations. This effect was most pronounced for anger, grievance, and negativity, particularly in news contexts (for example, anger demonstrated substantial reinforcement across all categories: 138% higher utility in News, 271% in Fitness, 116% in Gaming, and 90% in Random). A key concern emerging from our study is the platform's tendency to perpetuate negative emotions (e.g., anger, grievance). We find recommendation algorithm's reinforcement can create feedback loops that keep users locked into negative emotional states. Beyond the immediate user experience — where a constant feed of aggrieved or outraged content risks affecting well-being—there are broader implications for public discourse. **H, Habib, R. Nithyanand, YouTube Recommendations Reinforce Negative Emotions, 2025**

Relevant metrics we expect all VLOPs to share (by Integrity Institute) (*)

Scale

- How many people are exposed to violating content? (Per week/month/etc.)
- What % of users are exposed to violating content?
- How many exposures to violating content occur each week/month/etc.?
- What is the prevalence of violating content? (violating views/all views)

Cause

- What % of violating exposures occur because platform recommended the content top user?
- What % of violating exposures occur because the user followed account due to a platform recommendation?
- What % of violating exposures come from accounts with a history of posting violating content?

- Where violating exposures occur (what % occurs in feed, search, DMs, etc.)?

Nature

- Per violation, what is the breakdown in exposures by age, gender, country, other relevant vulnerable demographic? (so, what % of exposures to teens, what % to men/women, what % hate exposures to vulnerable populations)

Design

- For recommendation systems, how does the prevalence of violating content change as a function of the ranking and recommender scores?
- What are the key predictions that the ranking systems use? How does the prevalence of violating content depend on the predicted scores for each of those?
- If your system is predicting "like" probability, how does prevalence of violating content depend on p(like) scores? Is violating content more common for high p(like) scores?

Operation

- When conducting prevalence estimates, what % of violating content exposures occur on content that was never moderated? (So like, compare the overall prevalence of violating content to the prevalence of violating content that ends up getting moderated.)
- What is the distribution of views that violating content gets at the time of removal?
- What is the distribution of time delays when violating content is moderated?
- What is the ROC curve for machine learning classifiers of violating content? (what % of violating content is caught when the precision of the classifier is at 90%?)

Governance

- What are the key topline metrics for the company?
- Does violating content contribute to these metrics? What steps are taken to ensure violating content is not contributing to them?
- How often do changes to the platform that increase those metrics also increase the prevalence of violating content?
- How does the company manage when a change to the platform increases the company topline metrics but also increases the prevalence of violating content?

(*) numbers and stats should be broken down by policy violation and by country

3st theme:
Age restrictions, parental controls and granular user settings as (superficial) risk mitigation strategies

In their Risk Assessment reports VLOPs claim, with confidence, that their products are not available to users younger than 13. While still accepting self-declarations from minors, companies reassure us that they have implemented "a range of measures to detect if a user has inaccurately stated their age", including behavioural profiling, which uses a range of signals to predict a user's age.

How well are they doing in preventing kids younger than 13 from accessing their services? According to internal reports revealed in US litigation and independent research conducted in Poland – not very well… Instead of closing their gates for youngest users, VLOPs focus

on marketing of new user controls, including parental controls and settings to influence what they see on their own feed (keyword filtering, sliders "show me more"/"show me less", etc.).

This mitigation strategy shifts the responsibility for mitigating the harms of a system designed to keep users "hooked" onto children and their parents, as if the harms experienced by children could be prevented with more (self) discipline. Based on independent research (including interviews with children) CSOs warn that "user choice" and "parental control" will remain superficial and ineffective as long as platforms interfaces and algorithmic systems are designed to maximise time spent, at the cost of user autonomy and safety.

The table below shows what VLOPs tell us in their Risk Assessment reports about mitigating risk of excessive/problematic platform use (1st column), what we – CSOs – see through the lens of independent research (2nd column), and what we would really like to hear from VLOPs (what statistics and metrics we should like to see) in order to get a full picture (3rd column):

Quote from VLOPs' Risk Assessment reports (2025)	**Evidence and/or expert arguments challenging the effectiveness of the mitigation strategy**
TikTok "TikTok uses a range of measures to detect if a user has inaccurately stated their age at account registration, including automated moderation, which uses a range of signals to predict a user's age, and reporting channels, where users can report a "suspected underage user," if they believe the user is too young to use TikTok or to use a certain feature (such as LIVE). If an account with a potentially mis-stated age is detected, moderators will review it to determine whether they believe the account holder is under 13 and if they determine this is the case, the account will be banned."	**Methods used by VLOPs to control age are not effective:** "The results are unambiguous: all tested services [incl. TikTok, YouTube and Instagram] rely on self-declaration mechanisms for age checks when an account is created. None have implemented robust age assurance, and where parental consent tools do exist (notably on YouTube and Fortnite), they can either be easily bypassed or they are applied only after the account has already been created. In practice, minors can thus access these services freely, while platforms remain noncompliant with provisions that legally oblige them to treat minors differently than other users." **Interface EU, 2025 report, Mind the Gap** Only in Poland over 500,000 children aged 7-12 (over 1/3 of this age group) use TikTok, and the average time spent using this app is over an hour a day. One in four children aged 7-12 (630,000 children) uses Facebook. One in six children aged 7-12 uses Instagram (370,000 children). 36% of this age group (860,000 people) uses Messenger, and 31% (750,000 people) uses WhatsApp. **Institute for Digital Citizenship, Internet of Kids report, 2025**

Meta

"In order to use Facebook, users must be at least 13 years old."

"Meta has processes to identify and verify users' ages on its platforms allowing Meta to provide age-appropriate experiences. Depending on the platform, Meta requires users seeking to change their age from under 18 to over 18 to verify their age through various verification options, such as uploading an identification (ID), and/or recording a video selfie"

YouTube

"We implement a wide range of measures (such as minimum age requirements, signals for estimating the age of users, "made for kids" content…) to address the risk of children accessing age-inappropriate content. Although this risk will never be eliminated, our measures result in a significantly lower residual risk profile."

YouTube raised the minimum age required to livestream from 13 to 16 years old.

"For users who are under 18, Take A Break and Bedtime reminders are turned "on" by default. These are aimed at reinforcing healthy screen time habits. In YouTube Kids, our built-in timer also lets parents limit screen time

Age estimation based on profiling (behavioural signals) is privacy intrusive and unreliable, at the same time:

Respect the principle of data minimisation by prioritising anonymous age assurance approaches wherever possible. Invasive profiling methods are unlawful and can be highly biased.

Do not encourage additional first-party collection of biometric data beyond the scope of the existing service.

5 Rights, 2026 report

One expert cites statements from Meta that as of March 24, 2025, only 0.38% of youth users "predicted to reside in the U.S. were enrolled in Supervision through Family Center on Instagram." Plaintiffs' experts cite internal statistics as to Facebook as well, finding that only 0.15% of minors are enrolled in parental supervision tools. Further documents reveal that Meta reported that just 0.0038% of Instagram users had adopted parental controls as of March 2025, but the document does not specify the percentage of parents or minors enrolled.

According to US plaintiffs, **these levels of adoption are by design.** Meta introduced friction by requiring those with parental responsibility and minors to proactively navigate several steps to opt into the insights and controls.

KGI, Measuring Risk…, 2026

Negative feedback:

"Short term both implicit and explicit negative signals are effective at reducing the amount of topic content but the FYP algorithm can often "relapse": many accounts which cease to express disinterest and begin watching topical videos again can see their feeds dominated by such videos. The most effective explicit signal – marking a video as 'Not Interested' – is unintuitively buried in the interface.

Sapieżyński, 2025 (unpublished)

The 'Not interested' button [is] commonly expected to serve as an escape hatch from the toxic rabbit hole (people use it to avoid problematic, personally discomforting, or unrewarding content). (...) [But] most participants said they did not think their "Not Interested" was effective or felt that its effectiveness was inconsistent (confirming results from earlier similar studies based on user interviews from 2023).

Hong et al., The TikTok Algorithm Is Good, But Is It Too Good?..., 2025

Users complain feedback is not granular enough. Users were often uncertain of how each of their interactions was reflected in the algorithm. (...)

We observe individuals being unable to remove unwanted videos or genres of content from their FYPs, despite the active resistance to view such

by telling kids when it's time to stop watching." "Supervised experience on YouTube is for parents who decide their tween or teen is ready to access YouTube through a supervised Google Account. Videos a child can watch depend on the content setting their parent selects when setting up a supervised experience." "YouTube employs automated detection systems to find signals on YouTube channels that indicate that the channel may be owned by a user under the age of 13. These systems rely on content signals to find such channels, which are then flagged for a team to review more closely when they appear owned by an underage user."	content, a phenomenon we refer to as algorithmic persistence. Once captured in a content bubble, users found it difficult to leave, spending long periods of time being recommended undesired content. While the keyword filtering function technically worked such that the set of specified keywords stopped appearing on their FYPs, they still saw content from the same genre, all with hashtag content that had not been blocked. **J. Vera, S. Ghosh, They've Over-Emphasized That One Search…, 2025** Resetting feed: "refresh" feature is not a genuine reset of the algorithm – it is a temporary relaxation and all user data is retained. Refreshing" the algorithm will not prevent it from re-training on their client and becoming biased again over time. **Griffiths et. al., 2024**
TikTok "TikTok offers tools and settings to empower users to influence what they see on their FYF. This includes tools (...) to filter out certain keywords or hashtags to stop seeing certain content (...). As of March 2025, 2,619,937 users had actively filtered hashtags/keywords in the EU. Since its Year 2 Risk Assessment (...) TikTok has launched its "Manage Topics" tool, which allows users to customize the content they see in their FYF based on their preferences	

for a specific number of topics."	
Update on Instagram's Safety features (5/9/24): "Updated to clarify that we restrict adults over 18 from starting private chats with teens they're not connected to on Instagram and Messenger." "Meta offers users the ability to control how much of certain types of content (including sensitive or low quality content) they see in their feeds." "Meta has built age-related safety measures through features that minors can access to further protect themselves. Minors can adjust their settings and curate the content they see."	While some restrictions exist – such as preventing adults from messaging minors unless they follow each other – our testing found that teen avatars could also still freely send message requests to adults. The algorithm routinely encouraged our avatar accounts to widen their network of contacts with unknown strangers, increasing potential privacy and safety risks. It is unclear what content has been prioritised and why. This lack of clarity leaves young users unaware of how content is being pushed to them and may present skewed views of what is acceptable or popular. **5Rights, 2025 study**

Relevant metrics we expect all VLOPs to share (by Integrity Institute) (*)

Scale

- How many instances are there per week/month of underage users removed from the platform? (Broken down by country)

Cause

- For users that are removed for being underage, what steps were taken to validate their age when they signed up?

Operation

- What is the distribution of account age when accounts are removed for being under age? (For how many days are underage people on the platform?)

Governance

- Are there topline metrics for the company that underage users contribute to?
- What steps are taken to ensure that underage users don't contribute to company topline metrics?

(*) stats and numbers should be broken down by country

Conclusions

There is broad recognition that the way online platforms are designed impacts how users, communities, and societies experience benefits and harms from those platforms. With this brief, once again, we want to draw your attention to design features impacting the privacy, safety, and security of users – including minors – on digital platforms and the pervasive use of engagement-based recommender systems, which can be regarded as a form of addictive design. While the design of online platforms may exacerbate risks faced by users of all ages, adolescents have unique vulnerabilities that may lead to more negative experiences online (see eg. Office of the Surgeon General, "Social Media and Youth Mental Health: The U.S. Surgeon General's Advisory [Internet]" p. 13.).

As shown by quotes from Risk Assessments collected in this brief, VLOPs are not yet ready to acknowledge that their own design choices are (probably) the most influential risk factors when it comes to minors' exposure to harmful / illegal / age-inappropriate content and their problematic (extensive) use of social media. For the same reasons, companies focus on mitigation strategies that do not interfere with their business model, with short-term user engagement trumping other values (user safety, long-term engagement, user value).

Serious public debate about factors contributing to systemic risks generated by VLOPs' services, not only to minors, and the most efficient measures to mitigate these risks, won't happen as long as companies don't reveal key metrics, statistics and test results (see 3rd column in the tables above). Our questions directed to VLOPs, on the basis of their own claims made in Risk Assessment reports, do not aim at confidential business information. What we expect to see is data on the impact of mitigation measures designed and implemented by (allegedly) the most tech savvy and wealthy companies in the world. Surely these measures have been tested and evaluated internally. And if not – it is high time that they are.

Faced with these barriers, we call on the European Parliament to stand for robust DSA enforcement by:

1. supporting actions taken by the Commission on a political level;
2. supporting CSOs and independent researchers in their attempts to keep VLOPs accountable by demanding that companies present evidence to back up claims made in their Risk Assessment reports;
3. proposing solutions in the upcoming Digital Fairness Act that can strengthen the protection of most vulnerable consumers, such as minors, on online platforms.

Echoing the report on the protection of minors adopted by the Parliament last year, we also call on the Commission to make full use of its powers under the DSA to investigate the effectiveness of:

- age controls used by Very Large Platforms;
- mitigation measures supposed to protect children from exposure to harmful, illegal or age-inappropriate content;
- mitigation measures meant to prevent excessive use of VLOPs services, with special focus on addictive features that are used not only by TikTok (as shown by recent preliminary findings by the Commission) but by all VLOPs offering services to minors.

As the main enforcer of the DSA the Commission should also clarify how much space for maneuver is left by the DSA to national legislation, in particular as far as obligatory age-verification and digital minimum age are concerned. As a civil society organisation working on both EU and national level, we would like to see harmonised EU policy on how VLOPs should perform age verification and what age should they set as minimum to use their services, before Member States move forward with their respective legislations.