

Major Platforms, Minor Users

ECAT Research Workshop 2025: Book of Abstracts

2026

The background of the lower half of the cover features a series of concentric circles of varying sizes and colors, ranging from light purple to bright yellow. These circles are arranged in a pattern that resembles a stylized globe or a network of nodes, with some circles overlapping others. The overall effect is a vibrant, abstract design that complements the title and the year.

This document is a publication by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The contents of this publication do not necessarily reflect the position or opinion of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Contact information

Name: Emilie Sundorph

Email: emilie.sundorph@ec.europa.eu

Joint Research Centre

<https://joint-research-centre.ec.europa.eu>

JRC146333

PDF ISBN 978-92-68-39017-7 doi:10.2760/2290076 KJ-01-26-164-EN-N

Luxembourg: Publications Office of the European Union, 2026

© European Union, 2026



The reuse policy of the European Commission documents is implemented by the Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Unless otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

For any use or reproduction of photos or other material that is not owned by the European Union permission must be sought directly from the copyright holders.

All photos were taken by an official photographer of the event.

How to cite this report: Major Platforms, Minor Users: ECAT Research Workshop 2025 - Book of Abstracts, Sundorph, E., Punie, Y. and Pena Fernandez, E.A. (editors), Publications Office of the European Union, Luxembourg, 2026, <https://data.europa.eu/doi/10.2760/2290076>, JRC146333.

CONTENTS

ABOUT ECAT	2
ABSTRACT	3
WORKSHOP SESSIONS	4
Youth's lives: Presentation by students at Colegio Internacional de Sevilla San Francisco de Paula	4
ECAT activities on the protection of minors online	4
A captive audience? Social media addiction and the features that make it hard for children to leave	5
Communities of pain: Use of social media by youth at risk of eating disorders and self-harm	6
Agentic AI: Risks and opportunities for young people	8
Online influence on young people: Gender-based challenges and inequalities	9
ABSTRACTS	13
Ad personalisation and transparency in mobile ecosystems - A European perspective	13
Algorithmic transparency: The EU Digital Services Act and young people's experiences of online platforms	16
Children's protection online: The importance of discursive power in systemic risk management under the DSA	19
Defining, measuring and identifying solutions for mitigating online harms facing children and adolescents	21
Driven into the darkness: Amnesty Tech 2023	24
European regulatory principles for a safe internet for children compared with the perspectives and experiences of children and youth with mental health difficulties	25
Evaluating content moderation and systemic risks through the DSA's transparency framework	27
From doomscrolling to brainrotting: How VLOPs undermine autonomy and civic discourse through dysfunctional engagement?	29
Investigating the harms of viewing online misogyny on young people	31
Studying minors and machines: rethinking longitudinal research on algorithmic influence	33
Systemic risks to minors on online platforms: evidence from three studies by Landesanstalt für Medien NRW	36
Transparency in practice? Platform moderation trends after Romania's annulled presidential election	38
Waking up to smartphones: Towards evidence based solutions	41
X's Community Notes: Algorithmic resolution of crowd-sourced moderation on X in polarised settings across countries	42
CONCLUSIONS	46
LIST OF FIGURES	48
LIST OF TABLES	49

ABOUT ECAT

The European Centre for Algorithmic Transparency (ECAT) was launched in April 2023, as part of the European Commission's Joint Research Centre (JRC). ECAT provides scientific and technical expertise to support the enforcement of the Digital Services Act (DSA) and research into the impact of algorithmic systems deployed by online platforms and search engines.

The team's mission is to contribute to the implementation and enforcement of digital legislation by providing scientific guidance in areas such as safeguarding democratic values, protecting minors online, and ensuring accountability in the design of algorithmic systems and related technologies.

ABSTRACT

In November 2025, ECAT hosted a Research Workshop in Seville, Spain. The purpose of the event was to bring together researchers from across Europe to share insights on systemic risks of online platforms, particularly of very large online platforms and search engines (VLOPs and VLOSEs), as set out in the Digital Services Act (DSA). For this edition of the Workshop, the specific focus was on potential risks to the mental and physical health of children and young people, and how these can be mitigated.

Preceding the workshop, ECAT ran a call for contributions, and of the 52 submissions, 14 were selected to join to share their insights in poster sessions or presentations. In addition, the programme consisted of three panel discussions, a keynote, a presentation by local students and an overview of relevant ECAT activities.

From the perspectives of young high school students who kicked off the day, to top researchers in varied scientific disciplines such as communication studies, computer science and psychology and members of civil society, it was clear that the impact of online platforms on minors is an issue that requires careful and nuanced attention. Access to technologies that evolve so rapidly make it hard to keep up for parents, teachers, users, policymakers and researchers alike.

The main purpose of this document is to share some of the many insights that were shared throughout the workshop. In the first section, summaries of the workshop sessions are presented. In the second section, the selected abstracts for the event's poster sessions are reproduced. Hopefully these tasters will encourage readers to explore the work of the day's presenters further, and ultimately consider possible ways to contribute to this crucial discussion.

WORKSHOP SESSIONS

The aim of the event was to shed light on the ways in which online platforms affect young people, showcasing research examining the issue from a multitude of perspectives. While many of the themes resonate with adult users, they pose a markedly greater danger to vulnerable groups - particularly children.

Throughout the sessions, speakers identified specific risks but also outlined practical strategies for mitigation, drawing on the roles of platforms, families, schools and regulators.

Following a welcome from E. Alberto Pena Fernández, Head of ECAT, the programme began with a presentation by local students. Key insights from each session are summarised below, including references to some of the research cited and recommendations by researchers.

Youth's lives: Presentation by students at Colegio Internacional de Sevilla San Francisco de Paula



Student presenters: Andrés Almagro González, Artur Bonada Gómez, Carmen Escobosa Fernández, Inmaculada Concepción López Espejo, Qi Lin (Lucía), Rocío Pena Picón, Yago Taboada Viguera

The workshop kicked off with a presentation prepared by students at Colegio Internacional de Sevilla San Francisco de Paula, who had dedicated significant time and effort to conduct their own survey study of almost 100 fellow students.

Based on these insights, they opened the event with thoughtful reflections on the amount of time young people spend online, which was over two hours a day for more than half of survey respondents.

They also looked at which platforms were used the most, which was Instagram or TikTok for around half. Concerningly, despite respondents being minors, Pornhub was the platform most frequently visited by 6% of the respondents. While only a quarter of respondents said they felt influenced by social media, the student researchers reflected critically on this assertion, hypothesising that social media users may not always be aware of the way the platforms affect them. After offering up both the potential positives and negatives of social media use, the presentation concluded by suggesting that the two most common issues of platform use were cyberbullying and addiction. The key solutions proposed were for social media platforms to implement more effective content moderation in response to reports, and to remove any financial incentives of creating addictive platforms.

ECAT activities on the protection of minors online



Before heading into the event's first researcher panel, Dr Yves Punie, Deputy Head of ECAT, shared an overview of some of ECAT's own research and policy support activities. After first outlining the importance of the protection of minors online for the European Commission, he shared insights into publications and platform investigations related to the day's main topic.

In 2024, ECAT published an umbrella review of the evidence on the impact of social media use on the wellbeing of adolescents (Sala, et al., 2024). It found mixed results and highlighted both positive and negative potential. The study also emphasised that there are several intervening factors when it comes to mental health outcomes, relating to the characteristics of the individual, how they use social media, and how the platform functions.

From this mapping of the scientific literature, ECAT engaged in conversation with relevant researchers in the context of a series of roundtables (Manolios, et al., 2025) and provided scientific evidence to the guidelines on the protection of minors under the DSA published by the European Commission in July 2025 (European Commission: Directorate-General for Communications Networks, 2025). These guidelines set out a non-exhaustive list of proportionate and appropriate measures to protect children from online risks such as grooming, harmful content, problematic and addictive behaviours, as well as cyberbullying and harmful commercial practices.

ECAT also provides scientific and technical insights for ongoing cases under the DSA. For example, investigations are currently open relating to rabbit holes and addiction on platforms such as TikTok and Instagram, in addition to investigations on age verification or age assurance on those same platforms, as well as porn platforms. Overall, ECAT seeks to provide a strong evidence base and to balance risks and opportunities across scientific and technical support.

A captive audience? Social media addiction and the features that make it hard for children to leave



CHAIR

Astrid Bertrand, Project Officer, ECAT

PANELLIST

Susanne Baumgartner

is an Associate Professor at the University of Amsterdam. She is a member of the Center for Research on Children, Adolescents, and the Media, and the Digital Communication Methods Lab. Her research focuses on the role of digital media in adolescent development. Specifically, she is interested in how digital technologies affect the cognitive and emotional well-being of youth. In addition, she studies how social media design features affect adolescents' social media engagement and social media overuse. She employs innovative methodological approaches, such as experience sampling, smartphone tracking data, and longitudinal field studies. She received several rewards and grants, among others from the International Communication Association and from the Dutch Research Council.

Elisa Benedetti

is Researcher at the Laboratory of Epidemiology, Institute of Clinical Physiology - National Research Council. Dr. Benedetti's formal training includes Political Science and International Relations and a PhD in Economics. Her research focuses on monitoring of substance use and addictive behaviours in the general and student populations, as well as evaluation of the impact of policies on supply and demand for addictive goods in drug and gambling markets. She is the project manager of the European School Survey Project on Alcohol and Other Drugs (ESPAD), the biggest research project on substance use among adolescents in Europe, and scientific advisor to the Council of Europe for the coordination of the Mediterranean School Survey on Alcohol and Other Drugs (MedSPAD).

Alberto Monge Roffarello

is an Assistant Professor at Politecnico di Torino, Italy, where he researches the intersection of Human-Computer Interaction and digital wellbeing. His work has advanced understanding of attention-capture and deceptive design patterns, strategies for digital self-control, and frameworks that guide ethical, wellbeing-oriented design. His research has been published in leading HCI venues such as ACM CHI and ACM TOCHI, and he contributes actively to international debates on digital wellbeing and deceptive design.

He is also a member of the European COST Action on Digital Mental Health for Young People, where he collaborates with scholars and practitioners to inform policy and practice. At Politecnico, Alberto teaches courses on user experience design and digital wellbeing, mentoring students to design technologies that empower rather than exploit.

PANEL THEMES

In this discussion, panellists explored what we know about social media addiction and the impact of extended digital media use on young people. They also explored which platform features can exacerbate addictive behaviours, as well as the types of interventions, by platforms, individuals or regulation, that may be most effective in preventing compulsive use.

Alberto Monge Roffarello's remarks focused on how platform design features impact users of all ages, in what he calls the "Attention Trap". He shared several examples of studies showing how platform design is successful in keeping users engaged for longer, with features such as infinite scrolling and autoplay keeping users hooked. However, these so-called "attention-capture patterns" (Roffarello, et al., 2023) are not necessarily in the interest of user wellbeing. He has explored how to design interfaces that respect user attention (Roffarello, et al., 2024) but emphasised the importance of a multipronged approach to overcome the attention trap; user tools, design strategies and effective regulation all need to come together in support of wellbeing.

Susanne Baumgartner and her team have been carrying out intervention design studies to find out if specific social media design features affect the engagement and wellbeing of users, focusing on university students aged 18-22. Looking at various interventions, the studies suggest that smartphone gray-scaling (Dekker & Baumgartner, 2023) and a non-personalised feed, specifically on TikTok, does reduce screen time and can increase some measures of digital wellbeing, such as perceived control and lower levels of procrastination (Dekker, et al., 2025). Conversely, removing notifications did not seem to have much effect on smartphone use or subjective wellbeing measures, while it did increase feelings of "FOMO" (fear of missing out) (Dekker, et al., 2024).

Lastly, Elisa Benedetti presented a perspective on young people's use of digital technology in the wider context of addiction, through insights gained in the 2024 results of the European School Survey

Project on Alcohol and Other Drugs (EUDA & ESPAD, 2025). This survey has been running for 30 years and looks into the habits and attitudes of students aged 15-16 across Europe. Homing in on potential issues related to online platforms, Elisa shared findings on gaming, social media use and gambling. The proportion of students who perceive their own behaviour to be potentially problematic have increased across all three activities, but are by far the highest when it comes to social media, where almost half of students show an at-risk social media use.

Across the panel, there was agreement that children and young people are at particular risk when it comes to overuse of online platforms, and systemic shifts are required for platforms to prioritise wellbeing over engagement metrics. The discussion underscored the urgency of addressing digital behaviours as a public health priority, particularly for adolescents.

Communities of pain: Use of social media by youth at risk of eating disorders and self-harm



CHAIR

Emilia Gómez, Team lead, ECAT

PANELLIST

Florian Arendt

is Associate Professor of Health Communication at the University of Vienna, Austria. His research primarily examines the role of media in the health domain, operating at the intersection of communication and the social sciences on the one hand, and medicine and public health on the other. Given the inherently interdisciplinary nature of health

communication research, he primarily employs theories and methodologies from communication science to contribute to these collaborative endeavors. He is the author of more than 100 scholarly articles, published in leading journals in communication science (e.g., *Communication Research*, *Journal of Communication*, *Health Communication*, *Journal of Health Communication*) as well as in medicine and public health (e.g., *British Medical Journal*, *Crisis*, *Journal of Clinical Psychiatry*, *Social Science & Medicine*).

Lotte Rubæk

holds an MSc in Psychology from the University of Copenhagen, is an authorised psychologist and a specialist in psychotherapy. She is leading a self-injury team in Child and Adolescent Mental Health Services (CAMHS) and she is the overall leader of a clinical academic group for self-injury in the Capital Region of Denmark. She has been a member of Meta's SSI global expert group for a number of years, and raised criticism when she left it. Lotte is currently serving as a witness in major U.S. court cases against the large tech companies. Throughout her career, Lotte has had non-suicidal self-injury (NSSI) as her primary focus area, and she has worked clinically with self-injuring and eating disordered adolescents since graduating in 2007. Lotte has published three books and several articles and book chapters about NSSI - including two chapters for the *Oxford Handbook of NSSI*. Lotte gives lectures throughout the country on self-injury and social media, among other things.

Diana Ramírez-Cifuentes

is a Senior Data Science Consultant and Researcher with a strong background spanning academia and industry. Her expertise covers social computing, natural language processing (NLP), generative AI for textual data, and information retrieval systems, with applications in mental health and broader sociotechnical challenges for social good. She investigates how online behaviours and linguistic patterns can reveal indicators of psychological distress, leveraging large-scale social media data to better understand mental health dynamics. Beyond mental health, her research explores NLP applications in disinformation, social engineering detection, and hate speech analysis, contributing to safer and more ethical online environments. She has developed data-driven methods to detect and mitigate online risks and to reduce exposure to harmful content amplified by social media algorithms, particularly in the context of eating disorders. Diana holds a PhD in Computer

Science, an MSc in Data and Connected Systems, and an Engineering degree in Information Systems and Computer Science.

PANEL THEMES

In this session, panellists discussed how the online lives of vulnerable minors can affect them positively and negatively, how risks can be effectively mitigated, and also whether at-risk youth may be identified and offered help through social media.

Florian Arendt started off the discussion by sharing insights from his research into the impact of the media on those at risk of suicide. He described it as a “double-edged sword”, with the potential to do harm in cases where sensationalist depictions of suicide correlate with increased suicide rates. However, media also offers the potential to do good, with indications that stories of hope, recovery, and access to resources reduce suicidality.

Arendt shared findings from two specific studies related to social media. The first study shows that exposure to self-harm on Instagram was associated with suicidal ideation, self-harm and emotional disturbance (Arendt, et al., 2019). The second study offers some hope, indicating that being exposed to content from an influencer with lived experience of a suicide attempt, which focused on hope, healing and recovery, elicited a reduction in suicidal thoughts and increased intentions of seeking help (Arendt, et al., 2025).

Lotte Rubæk presented insights and evidence from her clinical experience. She started off by highlighting that all types of self-harm have increased significantly in the past decade, and that minors are particularly vulnerable. She then laid out the ways social media may have contributed to this increase, in what she broadly categorised as indirect, direct and systemic risk factors. Indirect factors include the broader negative consequences of social media use among minors, which constitute known risk factors for self-harm, such as addictive use and the development of a negative body image.

Direct factors consisted of risks such as being exposed to a sense of “competition” when it comes to the severity of self-harm and getting the idea that self-harm is the way to access care in the community. The systemic risk factors were aspects such as platform content moderation and design, where Rubæk in particular highlighted an investigation from 2024 which found severe limitations in self-harm content removal, and recommendations that

led vulnerable youth to self-harm profiles (Digitalt Ansvar, 2024).

Lastly, Diana Ramírez-Cifuentes shared research findings from the STOP project. The project aims to characterise how signs and symptoms of mental disorders are manifested on social media, to define models that can assess the mental state of users at risk, and non-intrusive ways of providing support through social recommendations.

Her presentation focused specifically on experiments related to social media users at risk of anorexia nervosa, a condition that strongly affects young people, especially girls. In a recent study, her and colleagues worked with clinicians to link social media behaviour to signs of risk, using insights to build explainable, behaviour-driven detection models (Ramírez-Cifuentes, et al., 2025).

They also developed an alternative content recommendation model, which showed that users displaying signs of anorexia nervosa were willing to follow accounts with pro-recovery content, and that applying the model increased harmless recommendations by 55%.

These insights provided food for thought in a discussion that highlighted the potential harm caused by recommender algorithms and showed that alternatives can be developed. Across the panellists there was agreement on the responsibility of platforms to protect minors from detailed, vivid depictions of self-harm or suicide methods, and favouring responsible and preventative content.

Agentic AI: Risks and opportunities for young people



KEYNOTE SPEAKER

Nicole Krämer

is Full Professor of Social Psychology, Media and Communication at the University of Duisburg-Essen, Germany, and Director of the Research Center "Trustworthy Data Science and Security". She completed her PhD in Psychology at the University of Cologne in 2001. Dr. Krämer has been pioneering interdisciplinary research on human-technology-interaction and opinion building in social media for 20 years. Recently, she increased research efforts in how vulnerable groups such as children communicate with AI. She served as Editor-in-Chief of the Journal of Media Psychology and currently is Associate Editor of the Journal of Computer Mediated Communication.

KEYNOTE THEMES

The keynote focused on the risks and opportunities of generative AI for young people, particularly in the context of anthropomorphisation, information processing, and relationship-building. Krämer highlighted how AI systems are designed with human-like features, indicated through typing behaviour and personalised dialogue, which can foster trust with users, and even make them perceive the system as benevolent in cases where it is not.

Younger children aged 5–12 are especially prone to anthropomorphising AI, and combined with a limited understanding of privacy risks, they may inadvertently disclose highly personal information to agentic AI systems. This kind of personal information may also be disclosed as part of teenagers' frequent discussions of personal issues with chatbots, which are perceived as non-judgemental.

She highlighted that while this type of interaction may provide some emotional support, it also comes with a risk of harmful dependencies. Another risk comes through the increasing use of AI instead of search engines for tasks such as homework, which may lead to reliance on greater rates of false or made-up information and undermine media literacy.

Taken together, she highlighted that while promising initiatives are underway both in research and regulation, more is still needed to gather solid empirical evidence on the impact of agentic AI on minors, and to ensure tech companies cannot conduct harmful "field experiments" on the population at large. Krämer strongly emphasised the need

for education initiatives from a very young age, to make children aware of the nature of AI and its associated risks.

Online influence on young people: Gender-based challenges and inequalities



CHAIR

Emilie Sundorph, Project Officer, ECAT

PANELLISTS

Harriet Over

is a developmental and social psychologist based at the University of York. Before moving to York, she worked at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany. She completed her PhD at Cardiff University in 2010. Her research explores the origins of prejudice and discrimination in childhood. Her recent work bridges psychology, digital culture, and mental health with a particular focus on online harms. She is especially interested in how online misogyny is affecting the behaviour and experiences of young people. Harriet's work has been supported by the European Research Council, Smart Data Research UK, the Economic and Social Research Council, and the Leverhulme Trust. She is principal investigator of the new HATESHIELD project, an EU-funded initiative to develop research-led interventions that encourage more egalitarian gender attitudes among children and young people.

Nicola Righetti

is a social and communication scientist specialising in quantitative and computational methods for the study of digital society, media, and public communication. His research advanc-

es methodological approaches and examines how digital media platforms shape communication, public discourse, and political behaviour through rigorous, data-driven inquiry grounded in the social sciences. He has published extensively on backlash against gender equality and its diffusion through social media networks.

Craig Haslop

is Senior Lecturer in Media in the Department of Communication and Media at the University of Liverpool, United Kingdom. Dr Haslop has published extensively about contemporary masculinities on social media and television. He is Principal Investigator for projects funded by the Office for Students and Economic and Social Research Council, which explore issues relating to masculinities online and development of educational interventions to raise young men's critical awareness of gender and sexual based violence. His recent research focuses on boys and young men's relationships with influencers and the mainstreaming of misogynistic discourses from the manosphere. He is co-author of the recently launched #Men4Change toolkit - a workshop resource designed to be used by professionals to help raise young men's (aged 18-25) critical awareness of harmful gender norms online and offline.

PANEL THEMES

The discussion focused on the emergence and popularisation of the so-called manosphere, broadly defined as online communities of anti-feminist and misogynistic content. Panel experts discussed the role of social media in mainstreaming extreme gender ideologies and what makes young people particularly vulnerable to this type of content.

Harriet Over focused on a recent research project exploring how the consumption of misogynistic content influences the behaviour and experiences of young people (Over, et al., 2025). Surveying 200 teachers in the UK, it found that 44% of secondary teachers strongly agreed that they were concerned about the impact of online misogyny on their pupils.

Perhaps more surprisingly, 24% of primary school teachers also strongly agreed with this statement. Through qualitative research, several examples were identified of pupils in both primary and secondary school explicitly referring to "manfluencers" (manosphere influencers) as justifications for disrespectful behaviour. In a second study with 400 teachers, Over also found that higher levels of

positive mentions of misogynistic influencers were correlated with higher rates of incidents of disrespect by male pupils towards both female pupils and teachers. In a third study, they discovered that greater popularity of “manfluencers” correlates with poorer mental health of female teachers.

Nicola Righetti presented findings from his research on the mainstreaming of anti-feminism. Drawing on survey data, he showed that anti-feminist sentiments are no longer confined to fringe online communities but are increasingly normalised across society. While social media influencers play a role, politicians and political parties contribute by appropriating anti-feminist tropes to increase visibility among like-minded communities. He also illustrated specific social media strategies that enable the spread of anti-feminism. In some political communities—especially on the far right—content invoking anger achieved greater reach, revealing a mechanism that facilitates the diffusion of anti-feminist narratives.

Righetti further demonstrated how coordinated networks of accounts, often linked to advocacy groups, exploit platform algorithms to amplify such content on social media. Beyond identifying mechanisms of digital diffusion, he highlighted data blind spots that hinder a full understanding of the spread of digital anti-feminism and advocated for greater social media data access for researchers.

Craig Haslop has run several focus groups with boys aged 13-14 to uncover their relationship with the manosphere, including specific influencers such as Andrew Tate. A study based on a 2022 focus group found that the boys’ interest in Tate came down to several different factors, including that they found him funny and an authentic voice which spoke up for them (Haslop, et al., 2024).

While one of Haslop’s recent studies outlines the waning of Tate’s influence (Haslop & Ringrose, 2025), this does not spell the end of the manosphere or other manfluencers like Tate, with data suggesting that sexist attitudes and behaviours are still on the rise among boys.

Reviews of other manfluencers’ accounts show that misogyny is often mixed with other popular content, especially related to fitness, finance, gaming and dating. Throughout, narrow ideals of masculinity, anti-feminism and often subtle messaging which blames women for men’s hardships are being promoted.

The overarching messages emerging from the session focused on the importance of engagement with children and young people. This needs to include specific efforts to dispel sexist myths and highlight positive role models, but also more generally to think critically about social media and the motivations of influencers.

Furthermore, it was highlighted that to truly understand how the content spreads and who it reaches successfully, researchers need greater access to anonymised platform data related to demographics and location, among others.

BIBLIOGRAPHY

Arendt, F., Scherr, S. & Romer, D., 'Effects of exposure to self-harm on social media: Evidence from a two-wave panel study among young adults', *New Media & Society*, Vol. 21, Issue 11-12, 2019, <https://doi.org/10.1177/1461444819850106>.

Arendt, F., Till, B., Gutsch, A. & Niederkrotenthaler, T., 'Social media influencers and the Papageno effect: Experimental evidence for the suicide-preventive impact of social media posts on hope, healing, and recovery', *Social Science & Medicine*, 2025, <https://doi.org/10.1016/j.socscimed.2025.117852>.

Dekker, C. A. & Baumgartner, S. E., 'Is life brighter when your phone is not? The efficacy of a grayscale smartphone intervention addressing digital well-being', *Mobile Media & Communication*, Vol. 12, Issue 3, 2023, <https://doi.org/10.1177/20501579231212062>.

Dekker, C. A., Baumgartner, S. E. & Sumter, S. R., 'For you vs. for everyone: The effectiveness of algorithmic personalization in driving social media engagement', *Telematics and Informatics*, Vol. 101, 2025, <https://doi.org/10.1177/20501579231212062>.

Dekker, C. A., Baumgartner, S. E., Sumter, S. R. & Ohme, J., 'Beyond the Buzz: Investigating the Effects of a Notification-Disabling Intervention on Smartphone Behavior and Digital Well-Being', *Media Psychology*, Vol. 28, Issue 1, 2025, pp. 162-188, <https://doi.org/10.1080/15213269.2024.2334025>.

Digitalt Ansvar, 'InstaHARM: En undersøgelse af Instagrams manglende indholdsmoderation af selvskadeindhold', November 2024.

ESPAD Group, ESPAD Report 2024: Results from the European School Survey Project on Alcohol and Other Drugs, EUDA Joint Publications, Publications Office of the European Union, Luxembourg, 2025.

European Commission: Directorate-General for Communications Networks, Guidelines on measures to ensure a high level of privacy, safety and security for minors online, pursuant to Article 28(4) of Regulation (EU) 2022/2065, Official Journal of the European Union, Luxembourg, 2025, <http://data.europa.eu/eli/C/2025/5519/oj>.

Haslop, C. & Ringrose, J., 'Post-Tate, post-truth, post-digital: researching and mitigating the misogyny influencers' *Gender and Education*, 2025,

pp. 1-17, <https://doi.org/10.1080/09540253.2025.2568408>.

Haslop, C., Ringrose, J., Cambazoglu, I. & Milne, B., 'Mainstreaming the Manosphere's Misogyny Through Affective Homosocial Currencies: Exploring How Teen Boys Navigate the Andrew Tate Effect', *Social Media + Society*, Vol. 10, Issue 1, 2024, <https://doi.org/10.1177/20563051241228811>.

Manolios, S., Sala A., Sundorph, E., Chaudron, S., Gomez, E. (eds), *Minors' health and social media: an interdisciplinary scientific perspective*, Luxembourg: Publications Office of the European Union, 2025, <https://dx.doi.org/10.2760/3795891>.

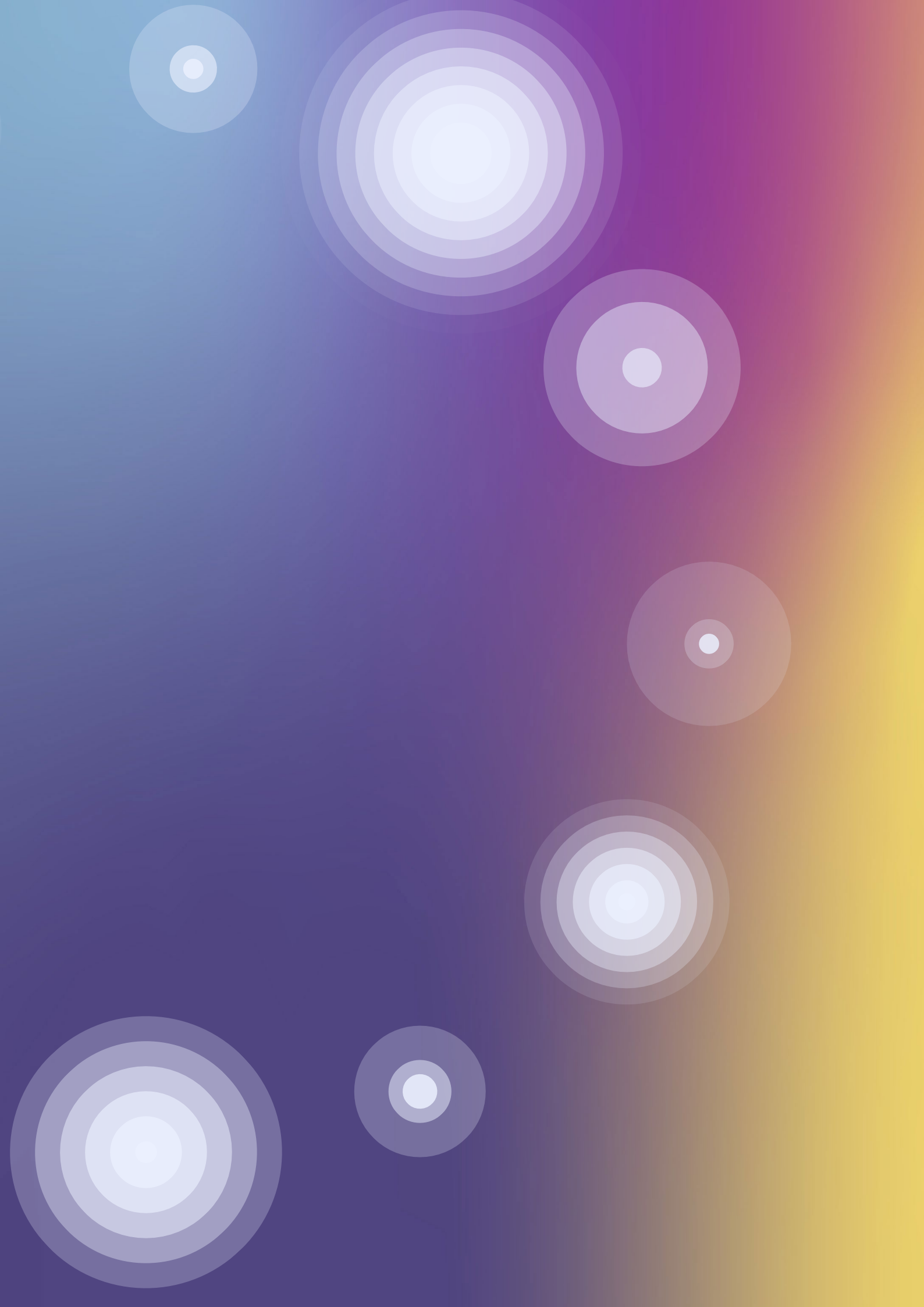
Over, H., Bunce, C., Baggaley, J. & Zendle, D., 'Understanding the influence of online misogyny in schools from the perspective of teachers', *PLoS ONE*, Vol. 20, Issue 2, 2025, <https://doi.org/10.1371/journal.pone.0299339>.

Ramírez-Cifuentes, D., Baeza-Yates, R., Lozano, M. & Freire, A., 'Enhancing contact recommendation in social platforms through mental health awareness: Exploring Anorexia Nervosa as a case study', *PLoS ONE*, Vol. 20, Issue 2, 2025, <https://doi.org/10.1371/journal.pone.0312766>.

Roffarello, A. M., Lukoff, K. & Russis, L. D., 'Defining and Identifying Attention Capture Deceptive Designs in Digital Interfaces', *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1-19, <https://doi.org/10.1145/3544548.3580729>.

Roffarello, A. M., Russis, L. D. & Pellegrino, M., 'Digital Wellbeing Lens: Design Interfaces That Respect User Attention', *AVI '24: Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, 2024, pp. 1-5, <https://doi.org/10.1145/3656650.3656674>.

Sala, A., Porcaro, L. & Gómez, E., 'Social Media Use and adolescents' mental health and well-being: An umbrella review', *Computers in Human Behavior Reports*, Vol. 14, 2024, <https://doi.org/10.1016/j.chbr.2024.100404>.



ABSTRACTS

In advance of the workshop, ECAT ran an open call for contributions, inviting researchers to submit abstracts to be considered for the poster session and short presentations¹. Most of the accepted submissions relate to the event's main topic, the potential risks of online platforms to minors, but a few were also of wider interest to the identification and mitigation of systemic risks. The selection of submissions was conducted by members of the ECAT team and based on relevance to the workshop topics, clarity, robustness of methodology, maturity of the work, and diversity (in terms of approaches, disciplines and types of VLOPs/VLOSEs covered).

The accepted abstracts are reproduced below, in alphabetical order of the abstract titles.

Ad personalisation and transparency in mobile ecosystems – A European perspective

David Breuer (TU Darmstadt), Lucas Becker (TU Darmstadt), Matthias Hollick (TU Darmstadt)

Note: This is an extended abstract of the corresponding paper by David Breuer, Lucas Becker, and Matthias Hollick: "Ad Personalization and Transparency in Mobile Ecosystems: A Comparative Analysis of Google's and Apple's EU App Stores", Proceedings of Privacy Enhancing Technologies 2026 (Issue 1). <https://doi.org/10.56553/pets-2026-0031>. Figure 1 stems from this paper.

NOTE

¹ To find the details of the Call for Contributions, see the ECAT website: https://algorithmic-transparency.ec.europa.eu/ecat-research-workshop-2025-call-contributions_en

Introduction

Consumers' choice is limited to the following two default options for installing smartphone apps: Apple's App Store for iOS or Google Play for Android. These stores are closely connected to the respective mobile operating systems. While third-party apps are subject to the platform's tracking rules and sandboxing systems, the app stores themselves do not necessarily follow those rules. Apple and Google use their stores to deliver personalised advertisements and recommendations to their customers, thus gathering user data to create targeting profiles. While they provide privacy policies stating which data they use, the exact functionalities of their targeting algorithms are unknown.

As such systems might entail and foster systemic risks according to Article 34 DSA, a thorough examination of such systems benefits society. To close this research gap, we examine the ad and recommendation personalisation behaviour of Apple's App Store and Google Play by conducting a large-scale sock puppet study on their smartphone platforms. In our work, we are interested in auditing the platforms in terms of their risks including "the protection of public health and minors and serious negative consequences to the person's physical and mental wellbeing" (Article 34(1) point (d) DSA) posed by the algorithmic recommender and advertisement systems as outlined by Article 34(2) DSA.

Both platforms are designated VLOPs according to the DSA and must provide online advertisement repositories. We aimed to use these repositories as the baseline for our experiments. Unfortunately, we identified issues in Apple's and Google's implementations that render those repositories useless

for us and potentially other researchers. Apple’s advertisement repository does not state any information on the number of advertisement recipients, and Google’s advertisement database does not contain any ads shown within the Google Play app.

Study

We conduct an automated black-box information flow experiment on smartphones to inspect Apple’s App Store and Google Play. We set distinct usage patterns and account characteristics as input parameters to our experiments. We measure the received ads and recommendations as output, as they contain information on the platform’s opaque profiling algorithms.

In contrast to similar studies, we execute all experiments on real smartphones to ensure the validity of our results. The platforms under test are entangled with their respective operating systems and hardware. Hence, using emulators or directly testing the APIs of the app stores might lead to improper results. We are interested in the effects of usage patterns and account types (called personas) on the advertisement and recommender systems of the two app stores.

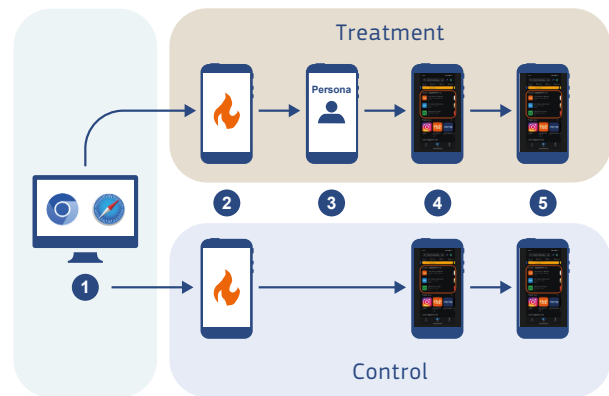
In our experiments, we use the following personas: Shopping, Finance, Parenting, Mental Health, Weight Loss, Alcohol Sobriety, and Gambling. These are signaled to the platform by installing distinct sets of apps through the particular store UI. Furthermore, we examine the effect of the following account parameters: Gender (inferred by first name), culturally significant names, and age. These parameters are set during account creation.

We conduct our experiment in measurement pairs consisting of control and treatment accounts (see Figure 1a). In Step 1, we create accounts for all our experiments. In Step 2, we reset and initiate our devices. This contains WiFi configuration and Apple or Google account login. In Step 3, we install a per-persona defined set of apps with the respective treatment account.

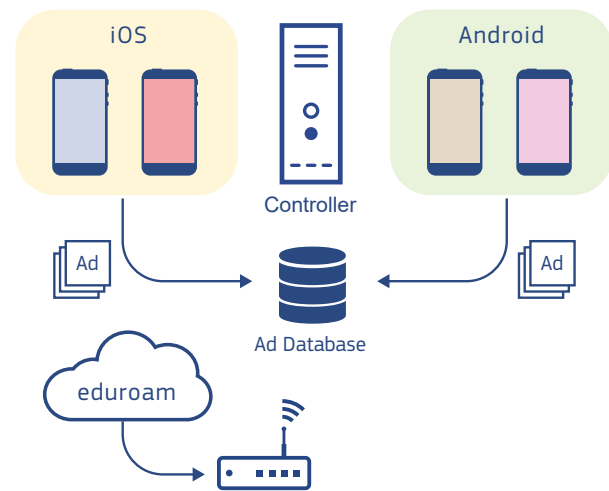
The control group does not install any apps. In Steps 4 and 5, we extract the received ads and recommendations on all devices before and after turning off personalisation of the app stores. We repeat each persona’s experiment five times per platform. We use two iPhones and two Android smartphones, which run the experiments per platform in parallel and fully remote-controlled by our experiment controller (see Figure 1b). The internet

traffic of the tested devices is routed via our university-wide WiFi network.

FIGURE 1
Overview of our experiment methodology and architecture



(a) Parallel measurement procedure of one corresponding control and treatment pair



(b) Overview of our experiment architecture instrumenting actual smartphones

Source: Authors' own elaboration

In total, we extracted over 1 million ads and 3.5 million recommendations using 362 Apple or Google accounts and 281 distinct eSIMs. Our study yields the following main observations:

(1) Google’s recommender system ignores user settings on sensitive ad categories. Users can exclude specific ad categories from their in-store experience. While Google does not show any ads on those topics, they are still shown as recommen-

dations. As those are hard to distinguish from a user's perspective, we view this as a potential risk according to Article 34 DSA.

(2) Apple's ad personalisation is subtle, while account parameters are more influential than user behavior, and personalisation is higher on the Search tab than on the Today tab of their App Store app. Apple's recommendations do not seem to be personalised at all.

(3) Very large ad campaigns dominate ads on both platforms. This means that the budget of the respective advertiser highly influences the ads shown to users. In contrast, we obtain clearer insights into Google's recommendations, which are not paid for, and therefore, advertisement budgets do not affect them.

(4) Some of our Google accounts were potentially flagged as child accounts. These accounts received no ads and could not access the Play Store pages for apps containing adult content (e.g., Dating).

While we do not know the exact behaviour that led to this automated decision, we see (a still unreliable) effort by Google to protect minors from ads. Nevertheless, Google still displays recommendations, which are easily confusable with ads.

Our results fluctuate while executing the same experiment multiple times. From this, we conclude that the tested systems exhibit a highly non-deterministic behavior. Unfortunately, this increases the difficulty of auditing such systems by black-box experiments.

Transparency shortcomings

While conducting our research, we identified several shortcomings in the transparency of the tested platforms regarding the DSA. Our initial plan was to use Apple's and Google's advertisement transparency repositories as the baseline data source for our experiments, but this is unfeasible due to the following shortcomings:

(1) Google does not provide any information on Play Store Ads as part of their Advertisement Transparency Center, as required by Article 39 DSA.

(2) Apple does not provide information on the number of recipients for individual ads in their Ad Repository as required by Article 39(2), point (g) DSA.

(3) Google does not mark contextual ads within the Play Store as Article 26(1), point (d) requires.

(4) Apple does not explicitly state the advertiser's name of an ad within the App Store as required by Article 26(1), point (b) DSA.

The usability of these ad repositories would be drastically improved if Apple and Google addressed the aforementioned issues, especially 1 and 2. We shared our findings with Apple and Google. While Google did not answer besides an automated reply, Apple discussed the individual remarks and justified their behaviour. Nevertheless, we still disagree with their interpretation of the DSA.

Hurdles analysing closed-source ecosystems

During our research, we encountered several hurdles inherent to the examined platforms that hinder transparency research on those platforms. We had to create hundreds of accounts on these platforms to analyse Apple's App Store and Google Play. As these platforms rely on valid phone numbers for account creation, we needed to acquire hundreds of eSIMs.

This increases the effort to conduct experiments on these platforms tremendously. Furthermore, we could not include certain app categories (e.g., dating apps) in our study as these require age verification to install them, which we could not easily bypass.

While this is an important measure to protect minors on these platforms, researchers should have an option to fully access those platforms from a user's perspective. Multiple of our accounts were banned or rate-limited during our research. This limited the number of ads and recommendations we could measure, resulting in a smaller dataset. We acknowledge the need for fraud detection systems on these platforms. Nevertheless, an option for researchers to audit the platform from a user's perspective, without the platform's knowledge, would benefit our community.

As these app stores are deeply integrated into their respective smartphone operating systems, we conducted our experiments on actual devices. While this limits the amount and execution speed of our experiments, we believe in the necessity of this approach to get authentic data, similar to actual user data, from those platforms.

Conclusions

In our work, we examine the advertisement and recommender systems of two platforms that consumers cannot avoid due to their duopoly of the app store market. Consequently, both are classified as gatekeepers by the European Commission according to the DMA and VLOPs according to the DSA. We present a methodology to conduct automated experiments on real smartphones to analyse smartphone-first platforms.

We conduct a large-scale study on Apple's App Store and Google Play and provide first insights into these platforms. We share our methodology and experiences auditing them from a user's perspective so researchers can build on that and examine similar platforms in the future. Our source code and dataset are available under <https://github.com/seemoo-lab/appstore-ad-tools>.

Acknowledgments

This work has been co-funded by the Federal Ministry of Education and Research of Germany in the project Open6GHub (grant number: 16KISK014) and the German Research Foundation (DFG) in the project CRUST (grant number: 503199853).

Algorithmic transparency: The EU Digital Services Act and young people's experiences of online platforms

Megan Nyhan (University College Dublin), Pranav Narula (University College Dublin), Izzy Fox (Dublin City University), Kevin Doherty (University College Dublin), Ruihai Dong (University College Dublin), Barry O'Sullivan (University College Cork), Josephine Griffith (University of Galway), Susan Leavy (University College Dublin)

The EU Digital Services Act (DSA) is currently being enforced across the EU, and several countries are introducing online safety regulations. As efforts to regulate digital spaces continue, so does evidence highlighting risks young people face from inappropriate content online, raising concerns about potential long-term psychological harm.

To explore how the experiences of young people on large online platforms, such as TikTok and Ins-

stagram, align with the focus of transparency obligations under the DSA, this paper provides details of a project that designed a series of stakeholder engagement studies. The project aims to demonstrate how engaging directly with young people on their experiences online is a vital measure to both inform the progress of the DSA and guide its implementation across the EU.

Introduction

The European Union introduced the Digital Services Act (DSA), partly to address such issues of online harm, requiring platforms to publish transparency reports on content moderation efforts to enhance accountability and protect vulnerable users. Additionally, social media platforms enforce content moderation to balance free expression with user safety, legal compliance, and platform integrity (West, 2018; Veglis, 2014). They aim to prevent the spread of harmful content, including hate speech, misinformation, violence, harassment, child exploitation, terrorism, and explicit material, which can lead to real-world harm (Scott et al., 2023; Carlin, 2024). Given these regulatory and platform-driven efforts to moderate content, it is essential to examine how young people experience and interact with algorithmic content recommendations on Very Large Online Platforms (VLOPs) to evaluate the appropriateness and effectiveness of their moderation approaches.

As part of a project called ARTAI (Assessing Risk for Trustworthy AI), we are conducting focus groups to explore young people's experiences with algorithmic content recommendations on VLOPs such as TikTok, Instagram, and X. These focus groups examine how these systems shape young people's online experiences, highlighting both risks and benefits. The focus group discussions examine key aspects of social media participation, including AI influence, algorithmic bias, and stereotyping, the impact of content recommendations, the experience of exposure to harmful content, patterns in their online behaviour (such as how they navigate platforms, respond to content, and engage with others), and potential design improvements to enhance safety and user experience.

In compliance with the DSA, many large online platforms have published transparency reports and content moderation policies. We have evaluated these to understand how platforms classify violating content, document content moderation practices, and report the removal of material that breaches their content moderation policies.

Preliminary findings suggest inconsistencies in policy definitions and transparency reporting, making cross-platform comparisons challenging. Given young people’s experiences of exposure to harmful content across multiple platforms, cross-platform comparison is essential to assess whether moderation efforts are consistently applied and effective in mitigating risks. The initial insights from undertaking the series of focus groups highlight the value in engaging with these users to gain an understanding of their interactions with algorithmically recommended content and the extent to which current moderation efforts align with their experiences.

DSA transparency reports and cross-platform evaluation

A transparency report published by a VLOP, such as TikTok, is a report that provides information on the content moderation efforts of a platform, ensuring accountability and compliance with DSA regulations. Facebook and Instagram release biannual reports covering six-month periods, from April to September and October to March. X, however, has exhibited inconsistency in its reporting schedule, with one report covering August to October 2023 and another spanning April to September 2024. Snapchat’s reporting is similarly irregular, with recent data covering January to June 2024. TikTok initially reported shorter durations, such as a single month in September 2023, before transitioning to quarterly and semiannual reporting. These inconsistencies in reporting frequency and coverage peri-

ods hinder meaningful cross-platform comparisons of content moderation practices. Furthermore, each platform defines and classifies content violations differently. For instance, policies addressing sexual content and nudity may be labelled as "Sexual Content", "Adult Nudity and Sexual Activity", or "NonConsensual Nudity", despite broadly covering overlapping issues. In our analysis, we found these inconsistencies significantly hindered our ability to make cross-platform comparisons and perform policy analysis.

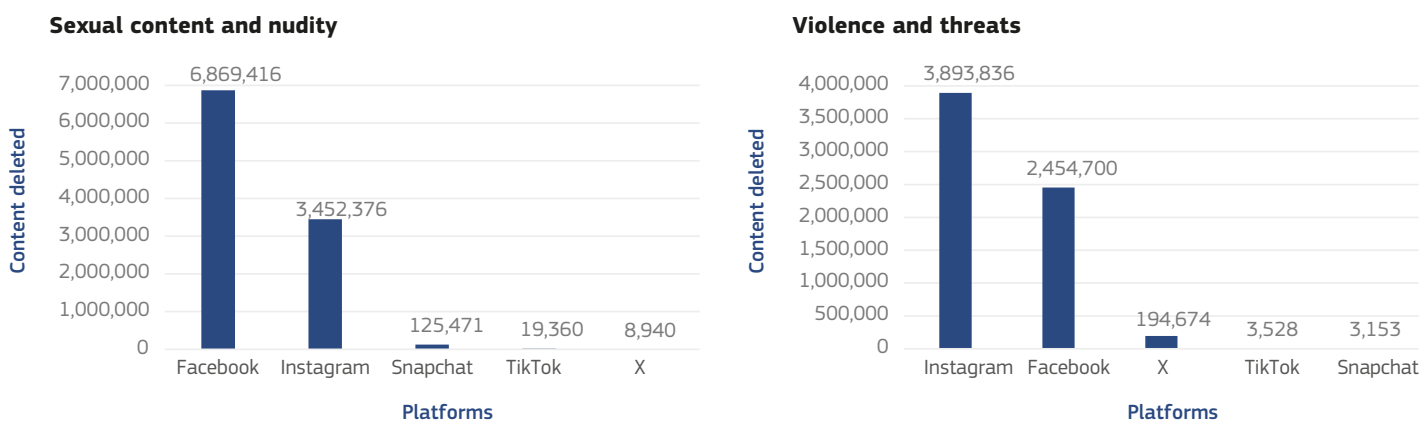
Without standardised reporting schedules, variations in data may reflect differences in reporting practices rather than actual enforcement effectiveness, making it difficult to track trends in harmful content over time. Additionally, irregular reporting allows platforms to control how and when they present data, potentially obscuring the true scale of harmful content and limiting regulatory oversight. Given young people’s reported exposure to harmful content across multiple platforms, these inconsistencies make it challenging to assess whether moderation efforts are effective in protecting vulnerable audiences.

To address these discrepancies and enable cross-platform comparison, we developed a standardised approach by aggregating content removal data from available transparency reports published between April 2023 and June 2024.

We matched different platforms’ policy violations, taken from their content moderation policies and transparency reports, to consistent categories.

FIGURE 2

Volume of content (including but not limited to reels, videos, static images and textual data) removed for 'sexual content and nudity' and 'violence and threats' across major social media platforms.



Source: Authors’ own elaboration, based on data from platform transparency reports

This made it easier to compare their moderation practices, even though they use different terms and reporting timelines. Since young people in our focus groups mainly discussed exposure to sexually explicit and violent content, we focused our analysis on these categories (see Fig. 2). Our findings reveal significant variation in content removal across platforms. For example, Facebook reports the highest volume of removals for sexual content violations, while Instagram leads in removing content related to violence and threats.

In contrast, X, TikTok, and Snapchat remove significantly less content across both categories. These disparities likely stem from differences in policy definitions, enforcement mechanisms, and platform-specific user behaviours. The variation in enforcement trends highlights the need for standardised reporting practices to support cross-platform comparisons and the broader assessment of content moderation effectiveness.

An alternative, for example, could involve mandatory, standardised quarterly reporting under the DSA framework, using a set of predefined categories with definitions and classifications of "harmful" content. This would help foster a landscape of transparency, enabling independent scrutiny and providing policymakers and researchers with reliable data to assess enforcement actions and their impact on platform governance.

Understanding patterns of algorithmic content dissemination

To evaluate the effectiveness of the DSA as it is implemented, it is crucial to understand the experience of end users of online platforms. Given the lack of aggregate data on how content is disseminated by recommender algorithms and levels of personalisation on online platforms, direct engagement with stakeholder groups is therefore crucial, but particularly challenging concerning young people. To address this, we developed a series of focus groups to engage with young people and understand the patterns of algorithmic content dissemination that they experience.

This can then provide insight into the effectiveness of the DSA as it is implemented. The study focuses on children within Irish secondary schools between the ages of 15 and 17. The primary objective of these focus groups is to develop a toolkit for engaging stakeholders in assessing the impact of recommender systems, particularly in address-

ing the ethical challenges when researching with young people.

This toolkit will serve as a structured framework for policymakers, researchers, and developers, providing best practices for ethical research, specifically when identifying risks and benefits of AI systems, and guidelines for improving transparency and accountability in platform design.

By engaging directly with users, we aim to uncover both the harms and benefits of these systems, potentially identifying previously unrecognised risks, including exposure to harmful content. Our focus groups use a think-pair-share format, engaging participants in discussions on five social media scenarios, including AI influence, algorithmic bias, harmful content, and platform design.

Facilitators then summarise key insights for further input. This method, known to enhance critical thinking (Kaddoura, 2023), encourages deeper reflection on the risks young users face online.

Flagging risk through stakeholder engagement

Preliminary findings from the series of stakeholder engagement studies being conducted highlight how the experience of young people can inform both how the DSA is enforced and future regulatory requirements. There was a notable gender dimension to the experiences of young people online, suggesting the prevalence of stereotyping in the kind of content that is recommended to them. For instance, many boys reported that their Instagram 'Explore' page was overwhelmingly populated with graphic car crash videos, where individuals were visibly injured or killed.

They also stated that they avoid using "X" (formerly Twitter) because their feeds frequently contain graphic videos of beheadings, shootings, and death. Some girls received similarly violent content involving shootings. Girls in the study were also recommended pornographic content on TikTok after viewing content related to fashion. They described how when they viewed content related to fashion, the recommender algorithm subsequently recommended sexually explicit content after a short period of scrolling. These reported experiences highlight concerns about content moderation and algorithmic transparency, which are central aspects of the DSA.

The exposure of young people to violent and sexually explicit content raises questions about the effectiveness of the platforms' moderation policies. Furthermore, engaging with young people in this way can highlight categories of content that should be included in DSA transparency reporting obligations.

Conclusion and future work

The goal of this project is to explore transparency measures under the DSA and evaluate how they align with the experience of young people on online platforms. We present an analysis of transparency reports and initial findings of a series of stakeholder engagement studies. Our research highlights inconsistencies in transparency reporting, content removal, and preliminary findings of stakeholder engagement with young people as users of online platforms, which highlight the experience of exposure to inappropriate material. Through this, we demonstrate how engagement with users of online platforms can provide insights into the effectiveness of the DSA and also indicate new categories of harmful content that should be included within transparency reporting obligations.

Acknowledgments

This work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224 and also with financial support of the EU Commission Recovery and Resilience Facility under the Science Foundation Ireland OurTech Challenge Grant Number 22/NCF/OT/11077 and 12/RC/2289_P2 at Insight the SFI Research Centre for Data Analytics at University College Dublin. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

Carlin, M., 'Real Harm to Real People: A Restorative Justice Theory for Social Media Accountability', *N. Ky. L. Rev.* 51, 2024, p. 145.

Kaddoura, M., 'Think pair share: A teaching learning strategy to enhance students' critical thinking', *Educational research quarterly* Vol. 36, Issue 4, 2013, pp. 3–24.

Scott, C.F., Marcu, G., Anderson, R.E., Newman, M.W., Schoenebeck, S., 'Trauma-Informed Social Media: Towards Solutions for Reducing and

Healing Online Harm', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23, 2023, pp. 1-20, <https://doi.org/10.1145/3544548.3581512>.

Veglis, A., 'Moderation Techniques for Social Media Content', in: Gabriele Meiselwitz (ed), *Social Computing and Social Media: 6th International Conference, SCSM 2014*, Springer International Publishing, 2014, pp. 137– 148, <https://doi.org/10.1007/978-3-319-07632-4>.

West, S.M., 'Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms', *New Media & Society*, Vol. 20, Issue 11, 2018, pp. 4366–4383, <https://doi.org/10.1177/1461444818773059>.

Children's protection online: The importance of discursive power in systemic risk management under the DSA

Elora Fernandes (KU Leuven), Andrea Palumbo (KU Leuven), Charlotte Ducuing (KU Leuven)

Problem statement

Under the DSA, systemic risk management is used as a regulatory tool to address political and policy questions around platform regulation. Given that such questions are framed in terms of systemic risks, the methodologies, terminology and approaches pertaining to risk management discursively shape the relevant political and policy questions. As a result, risk managers are conferred a form of discursive power to frame issues and set the agenda for systemic risk management (Griffin 2025).

While the DSA is essentially geared towards addressing the issues arising from the infrastructural power of Big Tech, this research explores the hypothesis that systemic risk management obligations may, paradoxically, enable providers of very large online platforms and search engines (VLOPSEs) to entrench their power and control the narrative. While this is a general statement that can be made for systemic risk management obligations, it carries specific consequences in the area of the protection of minors.

Rightly so, minors are granted a special status in the DSA as a vulnerable group. This is due to their increased susceptibility to online harms such as manipulation, exploitation, and harmful content, as well as their limited capacity to assess risks and exercise their rights (5Rights Foundation 2025; OECD 2021). In line with the principle of the best interests of the child (enshrined in both in Article 3(1) of Convention on the Rights of the Child and Article 24 of the Charter of Fundamental Rights of the European Union) the DSA imposes specific obligations on platforms to assess and mitigate systemic risks to minors (Recital 89).

This includes increasing transparency (Article 14(3)); putting in place appropriate and proportionate measures to ensure a high level of privacy, safety, and security of minors (Article 28(1)); prohibiting advertisement based on profiling of minors (Article 28(2)); or, more generally, specifically assessing the risk to minors (Article 34(1)(d)) and implementing targeted measures to protect them (Article 35(1)(j)).

However, focusing regulatory efforts on children (and other vulnerable groups) might be diverting attention from what really creates the power imbalances the DSA tries to tackle and undermining a more holistic understanding of systemic risks. Important features such as age-assurance mechanisms, content filters and parental control, while extremely important, should be seen as complementary measures to actually addressing the root of the problem: the economic incentives that drive the creation of opaque algorithms, the collection of vast amounts of data, and other engagement-maximising design patterns that generate risk and harm for all users (Asher Flynn, Zarina Vakhitova, Lisa Wheildon, Bridget Harris, Brady Robards, 2025).

Research objective and research question

This contribution aims to understand how the policy questions regarding the protection of minors are framed and understood by the actors involved in systemic risk management, with a particular focus on the agenda-setting role of VLOPSEs. It will map decision-making processes behind the assessment and mitigation of systemic risks to minors, which can provide us with insights on how these framings influence the prioritisation of systemic risks under the DSA, and to what extent an emphasis on protecting minors might shift atten-

tion away from the deeper structural drivers of harm.

Our analysis is conducted against the background of the literature on both the power of Big Tech and of the ensuing issues that they raise (see a.o. Gerbrandy 2023) and their specific risks to privacy, safety, and security of minors, especially their mental and physical health. This research will support the Commission in guiding and enforcing systemic risk requirements and especially in retaining the necessary discursive power for doing so.

Research methodology

This research combines doctrinal legal analysis with empirical desk research. The doctrinal and empirical methods complement each other to achieve the ultimate objective of understanding how policy issues regarding the protection of minors are framed.

Doctrinal research is conducted in two directions. First, on the relevant provisions that define how systemic risk management obligations are implemented, and how different actors participate in such implementation. These include not only Articles 34 and 35 of the DSA, but also all the provisions on supervision, enforcement and collaboration between regulated entities and the European Commission, and soft law instruments such as codes of practice and guidelines. The objective is to map the avenues for different actors to decide on how systemic risks to minors are assessed and mitigated.

Second, doctrinal research is carried out to identify the existing legal and regulatory guidance on the protection of minors on online platforms and search engines under EU and international law. By doing so, it is possible to determine which (if any) pre-existing guidance drives the decisions of relevant actors in assessing and mitigating systemic risks to minors.

The outcome sought with doctrinal research is to understand which actors are attributed the power to discursively frame issues around systemic risk management for the protection of minors, and how such power may be bound by existing guidance that limits their discretion. Furthermore, it provides insights into the possible prioritization by VLOPSEs of more 'symptomatic' risks over structural risks, that could be leading to symbolic compliance. This doctrinal research lays the groundwork for the empirical research.

The empirical component involves a desk analysis of publicly available materials, including transparency reports published by VLOPSEs under the DSA, decisions accessible in the DSA transparency database, policy documents and terms of service from VLOPSEs, and the documentation published by the European Commission, such as formal investigation notices and guidelines. The empirical research is carried out looking at decisionmaking patterns for all policy questions around the protection of minors. It also relies on process-tracing methodologies, to understand the processes behind certain policy approaches.

Selected references

5Rights Foundation, 'Risky by Design', 2025, <https://riskybydesign.5rightsfoundation.com>.

Flynn, A., Vakhitova, Z., Wheildon, L., Harris, B., Robards, B., 'Content Moderation and Community Standards: The Disconnect Between Policy and User Experiences Reporting Harmful and Offensive Content on Social Media', *Policy & Internet*, Vol. 17, 2025, <https://doi.org/10.1002/poi3.70006>.

Gerbrandy, A., 'Revisiting the Concept of Power in the Digital Era' in Andriychuk, O. (ed.), *Antitrust and the Bounds of Power On – 25 Years*, Bloomsbury Publishing, Oxford, 2023.

Griffin, R., 'Governing Platforms through Corporate Risk Management: The Politics of Systemic Risk in the Digital Services Act', *European Law Open*, Vol. 4, Issue 2, 2025, pp. 223–53, <https://doi.org/10.1017/elo.2025.17>.

OECD, 'Children in the Digital Environment. Revised Typology of Risks', *OECD Digital Economy Papers*, No. 302, OECD Publishing, Paris, 2021, <https://doi.org/10.1787/9b8f222e-en>.

Defining, measuring and identifying solutions for mitigating online harms facing children and adolescents

Sunny Xun Liu (Stanford University), Kang-Xing Jin (Stanford University), Anja Stevic (Stanford University), Angela Lee (Stanford University), Eleni Linos (Stanford University), Jeff Hancock (Stanford University)

The conversation around social media and youth wellbeing has grown increasingly complex, as researchers, policymakers, and industry leaders work to understand the true scope of potential harms and benefits. While new regulatory frameworks, such as the EU Digital Services Act (DSA), mandate risk-based approaches to assessing and mitigating online harms, many questions remain about how these risks are defined, measured, and addressed in practice. How do different harms impact young people? What strategies are most effective in reducing risks while preserving the positive aspects of social media use? Above all, how should we consistently measure the severity and probability of harms to inform learning about what is working and definitions of success?

To answer these questions, Stanford's Center for Digital Health and Social Media Lab gathered experts on mental health, youth psychology, technology, and policymaking to produce the 2025 Youth Safety and Digital Wellbeing Report. The report presents a harm taxonomy, harm measurement and mitigation strategies, the policy implications for defining, measuring and mitigating harms.

Harm taxonomy

Developing a shared understanding of the kinds of digital harms facing today's youth is a foundational first step in addressing and mitigating these threats. We reviewed four sources - the World Economic Forum's 'Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms' (2023); the National Academies' 'Social Media and Adolescent Health' report (2024); the Stanford Social Media Lab's 'Adolescents and Well-being Report' (2024); and the Minnesota Attorney General's 'Report on Emerging Technology and Its Effects on Youth Well-Being' (2024) - to understand dominant approaches to conceptualising and defining the various harms that youth may encounter online. These four sources were selected to provide coverage of possible social media harms from different perspectives.

We convened a group of 26 experts from academia, government, industry, and non-governmental organisations to discuss the categorisation of different harms using a modified Delphi method. Experts voted on whether each type of harm (e.g., "Exposure to unwanted violent or graphic imagery") should be included as a type of digital harm that should be addressed (1-9 Likert scale).

The group synchronously discussed each type of harm, shared their perspectives, and revoted. The

result of this iterative, collaborative process was to develop the Integrated Harm Framework (IHF), a list of 22 harms that online content may present to youth well-being and safety, with ratings and notes.

As shown in Table 1, each type of harm is categorised according to whether it 1) threatens

safety and/or relates to criminal activities (e.g., adult-minor solicitation, fraud), 2) harms health and well-being (e.g., exposure to communities that promote self-harm or suicide, increased symptoms of depression), 3) other content-driven harms (e.g., exposure to pornography in childhood, exposure to hate speech), or 4) other harms (e.g., undermining child privacy).

TABLE 1 (continues on the next page)
Integrated Harm Framework with final voting statistics and key points of discussion

Conceptual Categories:

-  1. Threats to safety and/or criminal activity
-  2. Health and wellbeing
-  3. Other content-driven harms
-  4. Other harms

	Specific Harm	Mean	SD	Mode	Key points of discussion
1	Adult-minor solicitation, image-based sexual abuse, sextortion	8.91	0.28	9	- Solicitation is among the most severe harms. - While related to 2, unlike CSAM the content alone is not illegal and necessitates distinct mitigation strategies, making separation from 2 advisable.
2	Child Sexual Abuse Material (CSAM) and Child Sexual Exploitation Material (CSEM)	8.87	0.45	9	- While not highly prevalent, this is among the most severe harms. - Suggestions to consider self-generated CSAM as a separate category.
3	Communities/content that promote self-harm or suicide	8.78	0.72	9	- High severity, though prevalence is low. - Rarity makes measurement difficult. - Recommendation to combine with 5, 15.
4	Bullying, harassment and stalking (including technology facilitated abuse and gender-based violence)	8.7	0.62	9	- Affects approximately 1 in 6 youth, although severity varies. - Distinctions between peer-to-peer bullying (which can originate in school) and other forms are important.
5	Communities/content that promote eating disorders, dysmorphia, unhealthy body image	8.52	0.88	9	- Highest-severity forms of this are relatively low prevalence. - Some harmful material can be disguised as wellness content. - Recommendation to combine with 5, 15.
6	Exposure to unwanted violent or graphic imagery	8.52	0.88	9	- Suggested as an additional category, or in combination with 9 under a broader category of "unwanted content".
7	Communities/content that promote terrorism or violence	8.35	0.91	9	- Concerns were raised about terminology. Replacing "promoting" with "facilitating recruitment for" was suggested.
8	Privacy (including unintended or unwanted exposure of personal information)	8.33	0.94	9	- User perspectives highlight privacy as a top priority. - Harm can stem from user misunderstandings of platforms' different settings. - Suggestion to separate concerns about i) platform data collection/use, and ii) exposure of information to other users.
9	Exposure to pornography in childhood	8.13	1.3	9	- Highly prevalent, affecting 1 in 12 children. Severity varies with the children as they age. - Suggestion to rephrase as "exposure to unwanted sexually explicit content in childhood". - Recommendation to combine with 6 under a broader category of "unwanted content".

10	Illegal transactions (including drugs)	7.83	1.27	9	- Limited data on prevalence, but self-reporting indicates use of social media for these purposes.
11	Loss of control, addiction, excessive use	7.74	1.92	9	- "Addiction" is a contested and possibly stigmatizing term. - Recommendation to combine with 13, 18.
12	Financial harms: underage access to illegal gambling or excessive spend on in-app transactions	7.57	1.38	9	- Limited data on prevalence, but anecdotal evidence of teenage access to online gambling via social media. - Not youth-specific; salience to youth users must be clearly articulated for risk assessment purposes.
13	Displacement of other beneficial activities (including sleep, exercise and in-person social activities)	7.48	1.78	9	- Mechanism, not a harm. - Recommendation to combine with 11, 18.
14	Misogyny, racism, hate speech	7.4	1.63	9	- Mechanism, not a harm. - Rooted in societal issues that predate and transcend social media.
15	Communities/content that promote dangerous challenges, or unsafe or unhealthy products	7.32	1.74	9	- Potentially very high-severity, including known youth deaths. - Recommendation to combine with 3, 5.
16	Fraud (including identity theft, impersonation, scams)	6.87	1.87	6	- Wide-ranging severity. - Not youth-specific; would require clear articulation of youth characteristics that could exacerbate general risk. - Exclusion from risk assessments could send the wrong signal.
17	Upward social comparison	6.26	2.19	9	- Mechanism, not a harm. - High prevalence. - Identified by youth users as a key issue.
18	Psychological impacts, including depression, sadness, anxiety, loneliness, lower positive well-being indicators, such as happiness, self-esteem	6.09	2.34	8	- High level of heterogeneity, with different hypothesized mechanisms, complicates inclusion in risk assessments. - Recommendation to combine with 11, 13.
19	Misinformation and disinformation	5.83	2.32	7	- Mechanism, not a harm. - Not youth-specific. - Rooted in societal issues beyond social media.
20	Infringement of Child Rights: over-limiting child access to information	5.74	2.22	5	- Suggested as an additional category following preliminary survey. - Not discussed in detail due to time constraints.
21	Algorithmic biases and risks (including recommendations of problematic content or discrimination in decision-making)	5.53	2.5	8	- Not well defined. - Mechanism, not a harm; should be considered in relation to other harms. - A tool that can exacerbate harms, but also underpins many mitigation strategies.
22	Parent use of social media and related stress/displacement of social interactions	5.27	2.56	5,9	- Mechanism, not a harm. - Worth addressing in parent-targeted materials.

Source: Authors' own elaboration

Our framework highlights several important considerations when defining and measuring harms. First, it is important to consider both the prevalence and severity of digital harms: some forms of harm, such as adult-minor solicitation may be

relatively rare but can have devastating consequences on child well-being, whereas other forms of harm such as overuse may be widespread but have less severe immediate effects.

Second, it is important to distinguish between mechanisms and outcomes when conceptualising digital harms. For example, social comparison processes are well-known to undermine adolescent well-being when they compare themselves to others, which results in the negative outcome of increased anxiety or worse body image. Clarifying how specific digital harms are the product of studied mechanisms can inform efforts to measure and mitigate them.

Third, the IHF highlights the importance of defining actionable harms and ensuring that they are specific enough and measurable enough to inform mitigation efforts, and evaluate the efficacy of such efforts in real time. Fourth, our discussions raised questions about youth-specific harms vs. general harms. Some harms are generally applicable to all age groups, whereas others may be differentially harmful to youth (e.g. 'exposure to unwanted pornography in childhood').

Some participants argued for more of a focus on the latter; others suggested prioritising by severity and prevalence for youth, regardless of whether something is uniquely harmful for youth. Finally, we discuss the relationship between societal issues and social-media-only issues. Some harms have pre-existing offline components (for example, bullying occurs in schools), and in some cases online components of those harms may only be part of broader societal issues. Some participants emphasized the importance of considering these harms as part of a broader system for both assessment (e.g., trying to understand relative prevalence and severity across channels) and mitigation (e.g., considering both online and offline resources in mitigations). Some also suggested a heightened priority for the harms that are differentially over-represented online.

Harm measurement and mitigation

The report presents three main approaches to measure harms: content or behaviour metrics, user feedback and reports, and validated offline outcomes. Content or behaviour metrics include screen time, frequency of interactions, and engagement with specific types of content, and random sample of content on platforms.

User feedback and reports metrics include representative user surveys, in-platform reports, and reports through external channels such as law enforcement. Validated offline outcomes metrics include data from other national surveys, such as the National Survey of Children's Health. The report fur-

ther presents five general mitigation strategies: Age certification, education, content moderation, algorithmic interventions and youth-interests-first design.

Our report further provides recommendations to platforms, policymakers, researchers to better assess and mitigate the harms and better support minors and families. Platforms shall enhance transparency and share data on prevalence, severity, mitigations with regulators and researchers. Platforms need to refine designs, such as remove addictive features, and prioritize quality over engagement. Policymakers shall establish consistent global standards to ensure proportionate safety measures across platforms, reducing disparities in harm mitigation efforts between countries and platforms while reflecting differences in organization size/capacity and risk surface. Policymakers need to create youth-specific guidelines, such as age-appropriate content moderation, data privacy, design features, and approaches to age verification. Researchers shall engage with youth, conduct practical and actionable research and bridge the gap between science, policy and the public.

Above all, there is an urgent need for policymakers, platforms, and researchers to work together to develop and implement standardized ways to measure the severity and probability of harms, as this is a foundational component that informs all of the above actions. We gathered a similar group of experts in a separate follow-up workshop to provide more specific recommendations on standardization of measures for risk assessment, with an additional report forthcoming shortly.

Driven into the darkness: Amnesty Tech 2023

Amnesty International

During the Covid-19 pandemic, TikTok emerged as a global platform, attracting hundreds of millions of children and young people largely thanks to its 'For You' page, an infinitely scrollable feed of personalised video suggestions, and the algorithmic recommender system behind it. In 2023, Amnesty International researchers in four countries together with partners at the Algorithmic Transparency Initiative and AI Forensics set out to understand the algorithmic pipeline behind harmful mental health-related content on the platform and its impact on young people. The findings were published in November 2023 in a report entitled "Driven into

the Darkness: How TikTok's 'For You' Feed Encourages Self-Harm and Suicidal Ideation".

We'll share insights from employing a mixed-methods approach to study the algorithmic recommender system on TikTok. We will delve into our approach which married a sock-puppet audit of the recommender system with in-depth interviews with children and young people centring the lived experiences and perspectives of young people in Global Majority countries.

In particular, we'll explore the key methodological challenges we encountered over the course of the project both on the quantitative and qualitative side of the research. Amnesty researchers found that through the recommender system's seamless hyper-personalisation, TikTok has created an addictive platform, despite mounting evidence of the serious health risks associated with children's compulsive use of social media.

Examining further risks of TikTok's content targeting, Amnesty International's research shows that TikTok's 'For You' feed can easily draw children and young people who signal an interest in mental health into "rabbit holes" of potentially harmful content, including videos that romanticise and encourage depressive thinking, self-harm and suicide. TikTok risks exacerbating children and young people's struggles with depression, anxiety and self-harm, putting young people's mental and physical health at risk.

Further research, released in 2025, brought together testimony from affected young people and parents in France as well as renewed technical research evidence. The briefing documented TikTok's failure to address its systemic design risks for children and young people including the company's binding obligations under the EU's Digital Services Act. It is an urgent appeal to the company itself, but also to EU and French regulators to take decisive action to force the company to respect children's and human rights.

Post-publication, Amnesty Tech is now campaigning and advocating to make TikTok safer, drawing also on EU transparency mechanisms including the first DSA risk assessment reports.

European regulatory principles for a safe internet for children compared with the perspectives and experiences of children and youth with mental health difficulties

Elisabeth Staksrud (University of Oslo), Mariya Stoilova (LSE), Sonia Livingstone (LSE), Richard Graham, Line Indrevoll Stänicke (University of Oslo), Reidar Schei Jessen (University of Oslo), Tine K. Jensen (University of Oslo)

This study examines the regulation of children's digital environments through parental mediation, industry self-regulation, and legal enforcement, as experienced by a unique set of children and young people from Norway and the UK, with known mental health difficulties. Through their experiences, key gaps and contradictions in existing frameworks are identified.

Findings indicate that parental awareness and control mechanisms are often ineffective due to low digital literacy, punitive approaches, and fear-based mediation, which discourage open communication. Simultaneously, industry self-regulation remains inconsistent, with inadequate content moderation and inaccessible reporting tools that fail to protect young users effectively and leave them in dismay. Legal enforcement is further complicated by contradictory guidelines and cross-border jurisdictional challenges, leading to low trust in law enforcement among young people.

The study argues that the trickle-down regulatory model, where responsibility cascades from policymakers to caregivers, is insufficient. Instead, an integrated, multi-stakeholder approach is required—one that prioritises trust-based parental mediation, stronger industry accountability, and clearer legal frameworks. These findings emphasise the need for dynamic, reciprocal regulation to ensure a safer digital environment for children.

This work was conducted as part of the ySKILLS project, supported by the European Union's Horizon 2020 Research & Innovation Programme under Grant Agreement no. 870612.

For further information about the overall study see Livingstone, S., Stoilova, M., Stänicke, L. I., Jessen, R. S., Graham, R., Staksrud, E., & Jensen, T.K. (2022). Young people experiencing internet-relat-

ed mental health difficulties: The benefits and risks of digital skills. An empirical study. KU Leuven, *YS-KILLS*.

Method

The corpus used in this study are interviews with young people with known mental health difficulties. We conducted interviews with 62 young individuals aged 12 to 22 (30 in Norway and 32 in the UK). The research was conducted according to relevant ethical guidelines and principles for Norway and UK respectively, and strict protocols for securing sensitive data were followed.

While the participants had varying degrees of mental health difficulties, most had received recent treatment. In the Norwegian group, all had been diagnosed with a mental illness, and interviews were conducted by clinical psychologists who were part of the research team.

In the UK, a psychiatrist was included in the interview team, ensuring support for the participants when talking about sensitive topics. Vigilances were in place towards detecting if there was suicide risk and contact with the health system was needed. For both teams, the informant's wellbeing was carefully monitored, and adjustments were made as needed.

The team also provided all interviewees the opportunity to discuss their feelings about being interviewed and potential adverse effects, and all were offered professional follow-up mental health support if needed.

Example of relevant findings

All informants mentioned social media companies and/or big tech companies in relation to their experiences and views on regulation. Through the analysis of the interview transcripts, many raise critical questions of implementation and actual practices of self-regulation, especially by social media companies.

A strong theme throughout the discussions is the disconnect between what social media companies claim and what users experience. Participants express frustration over platforms deflecting responsibility, with many noting that reporting mechanisms are ineffective.

Algorithms were seen to play a central role in shaping the digital landscape. Multiple participants recognised or postulated that platforms intentionally

design their algorithms to maximise engagement, often by promoting controversial or addictive content, raising ethical concerns about how self-regulation aligns with profit-driven motives rather than public safety. Based on the analysis of the interviews, the following dominant themes emerge:

(1) Perceived ineffectiveness of platform regulation: Social media platforms are seen as ineffective at moderating harmful content. Some interviewees believe platforms allow problematic content to thrive due to financial incentives, where platforms prioritising profit over safety, especially using algorithms to drive engagement even at the cost of user wellbeing, also pointed to not only an ineffectiveness, but a lack of corporate responsibility.

(2) Companies deflect responsibility, often claiming they are “just a platform” rather than taking accountability for the content they host and how unregulated content impact mental health. Participants describe significant mental health consequences resulting from exposure to harmful content, often exacerbated by a lack of moderation.

(3) Lack of enforcement: Interviewees express frustration with companies failing to enforce their own policies, especially regarding harmful content and community guidelines.

(4) “The illusion of control over content”: Informants emphasised that that while platforms claim to allow users control, the reality - in their actual experience - is that algorithms and policies dictate exposure to content. Participants note that harmful content, especially hate speech and mental health-related issues, is often ignored, while trivial content may be over-policed. Informants express that while social media companies claim to regulate content, their enforcement remains highly selective.

Implications and future directions

By bringing forward the voices and experiences of young people, this study provides empirical evidence that challenges dominant regulatory assumptions, moving beyond policy abstractions. It shows how young people perceive, navigate, and sometimes resist the very regulations meant to protect them.

In doing so, it contributes to ongoing discussions about how digital governance can be reimagined to better align with children's rights, agency, and everyday digital realities. In lieu of our findings, and in line with the promise to carry the unique

voices of these particularly vulnerable and experienced young people, we end the article with observations and recommendations for future policy and practice.

Evaluating content moderation and systemic risks through the DSA's transparency framework

Benedetta Tessa (University of Pisa/IIT-CNR), Amaury Trujillo (IIT-CNR), Stefano Cresci (IIT-CNR)

Introduction

This contribution summarises more than two years of quantitative research on the DSA Transparency Database and Transparency Reports related to the eight largest social media platforms in the EU. It summarises and reports key findings from a series of peer-reviewed studies published in prominent venues such as ACM CSCW 2025 (Trujillo et al., 2025), the COMPASS Workshop at AAAI ICWSM 2025 (Shahi et al., 2025), and the AIDEM Workshop at ECML-PKDD 2025 (Tessa et al., 2025).

In these studies we investigated platform compliance with the DSA's transparency obligations, with particular reference to their handling of the systemic risk (Art. 34) concerning negative effects on civic discourse and electoral processes around the 2024 European Parliament elections. Our results shed light on how the DSA is enforced in practice and allow identifying a set of changes needed to achieve greater transparency, accountability, and oversight. In detail, we revealed significant shortcomings in both the structure of the Transparency Database and Transparency Reports, and how they are utilised by the scrutinised platforms. Frequent issues include vague reporting and cross-source inconsistencies, especially when it comes to platform X (formerly Twitter).

Moreover, a detailed analysis of the Transparency Database revealed that no substantial shifts in moderation occurred during the EU 2024 electoral period, in spite of the heightened risks of electoral interference. Overall, these results expose technical and normative gaps that must be addressed for the DSA's transparency mechanisms to the deliver on their accountability promise.

Methodology

Our data-driven analyses focus on the statement of reasons (SoRs) submitted to the DSA Transparency Database (Art. 24(5)) and the corresponding Transparency Reports (Art. 15) of the eight largest social media platforms in the EU (i.e. Facebook, Instagram, LinkedIn, Pinterest, Snapchat, TikTok, X, YouTube).

We collected 1.93B SoRs across two distinct periods: (i) the initial 100 days after the database's launch in September 2023 (353M SoRs) and (ii) a later 100-days period from July to October 2024 (573M SoRs), about one year after the database's launch.

For the initial period, we assessed how platforms complied with the new obligations by examining SoR characteristics, delays between content creation and moderation, and the reported use of automation. We then cross-checked these findings with the platforms' Transparency Reports covering the same period, to assess consistency in reporting. The same analyses were carried out in the later period (i.e., one year later) to detect whether platforms adapted and improved their reporting after the initial settling phase.

Finally, we analysed SoRs from an extended period spanning from March to October 2024 (1.58B SoRs), surrounding the 2024 European Parliament elections, to detect possible changes in moderation and reporting practices around this politically sensitive period.

Main results

Our analysis of the first 100 days of the DSA Transparency Database revealed substantial shortcomings in both the consistency and reliability of the submitted data. To begin, we found notable differences among platforms in terms of restriction types, affected content categories, timeliness of moderation, and reliance on automation (Trujillo et al., 2025).

This result reflects the platforms' varying levels of adherence to the database's philosophy and different moderation approaches, indicating greatly heterogeneous moderation practices.

Moreover, consistency checks and comparisons with each platform's Transparency Reports uncovered widespread discrepancies, with platform X exhibiting the most inconsistencies between its submitted SoRs and the Transparency Report. As

a striking example, all SoRs submitted by X consistently reported no use of automation whatsoever, which starkly contradicts the platform's own Transparency Report (Trujillo et al., 2025). In addition, X's SoRs always reported no moderation delay, which is a highly implausible scenario given the lack of automation. Alas, after repeating the same analyses more than one year later, we found that the issues in X's SoRs persisted without improvement (Shahi et al., 2025).

Another widespread issue is the degree of uninformativeness of many of the SoRs in the database. All platforms formally complied with the DSA requirements by consistently filling all mandatory fields. However, most omitted optional yet crucial details, such as geographic scope or content language, thus undermining the database's utility for independent scrutiny. Across all platforms, only about 60% of optional fields were used at least once, and even then only sporadically. In addition, the vague and catch-all category scope of platform service consistently represents the most frequently reported infringement (41% overall in the later period, decreased only slightly from 42.7% in the initial period).

Therefore, our analyses brought to light persistent issues of vague and contradictory reporting, underscoring the need for processes capable of detecting inconsistencies and ensuring alignment across transparency mechanisms. A favourable step in this direction would be an automated process that compares the data submitted to the Transparency Database with the aggregations in the platforms' official Transparency Reports, allowing to streamline the checks that we have performed manually so far (Tessa et al., 2025). In this regard, the recent introduction of standardised templates for Transparency Reports can significantly facilitate this endeavour.

Negative effects on civic discourse and electoral processes

To specifically investigate possible systemic risks associated with the 2024 European Parliament elections held between 6 and 9 June 2024, we also carried out an analysis of the SoRs submitted to the Transparency Database from March to October 2024. Our analyses assessed whether platforms' moderation practices varied during this period, for instance due to the possible presence of election-related misinformation or foreign interference.

However, we found no meaningful shifts in moderation activity surrounding the electoral period compared to periods associated with lower systemic risk. Furthermore, we also analysed anomalies in moderation volume and delay that occurred during the electoral period to investigate their underlying causes. We compared anomalous moderation actions with those from routine days, analysing characteristics such as the type of content moderated, reasons for moderation, use of automation, etc. Overall, we found that the anomalies were unrelated to the elections, with the notable exception of LinkedIn, which took action in September against multiple election-related comments posted in June.

These results suggest that, overall, platforms may not have adjusted their moderation practices in response to the increased risks typical of large-scale political events. One possible explanation is that they deemed their existing moderation practices sufficient. An alternative explanation is that changes might have occurred but were not visible in the data. For example, platforms could have underreported or omitted relevant enforcement details, or the database's design, relying on predefined and often generic categories, may have been too limited to allow detecting meaningful shifts (Shahi et al., 2025). Regardless of the reason, our results reveal that, with the available tools, it is currently not possible to detect whether meaningful changes in platforms' moderation practices occur in times of heightened election-related risks.

Conclusions

Despite the promises of the Transparency Reports and the Transparency Database, our results show that these have not yet reached their intended potential due to persistent issues such as incomplete reporting, vague categorisations, and inconsistent data. These limitations significantly reduce the usefulness of such transparency mechanisms. For example, the current implementations do not allow to determine if and how platforms react to heightened systemic risks, such as those about election interference.

Our findings thus underscore the need to improve the quality and completeness of reporting, both from the platforms' and the regulators' side. On a positive note, though, ongoing efforts to harmonise Transparency Reports and the recent updates to the Transparency Database schema, including the introduction of more specific infringement categories and an increase in mandatory fields, represent desirable and needed improvements to-

wards reducing vagueness and un informativeness. The rollout of the dedicated data access platform pursuant Art. 40 of the DSA is further bound to enhance oversight and transparency by facilitating the implementation of verification processes to audit the trustworthiness and completeness of the SoRs submitted to the DSA Transparency Database, as envisioned in Tessa et al., 2025.

Nonetheless, it is still uncertain whether these improvements will be sufficient to overcome the serious and widespread issues we documented, and their actual impact will only become clear over time. Within this context of grave uncertainty, our studies provide regulators and policymakers with a rigorous and thorough quantitative assessment of the current shortcomings of the existing transparency mechanisms and their implications, while also offering guidance for future improvements.

References

Shahi, G. K., Tessa, B., Trujillo, A., & Cresci, S., 'A Year of the DSA Transparency Database: What it (Does Not) Reveal About Platform Moderation During the 2024 European Parliament Election', Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media, 2025, <https://doi.org/10.48550/arXiv.2504.06976>.

Tessa, B., Amram, D., Monreale, A., & Cresci, S., 'Improving regulatory oversight in Online Content Moderation' ECML-PKDD Workshops, 2025, <https://doi.org/10.48550/arXiv.2506.04145>.

Trujillo, A., Fagni, T., & Cresci, S., 'The DSA Transparency Database: Auditing self-reported moderation actions by social media', Proceedings of the ACM on Human-Computer Interaction, Vol.9, Issue 2, 2025, <https://doi.org/10.1145/3711085>.

From doomscrolling to brainrotting: How VLOPs undermine autonomy and civic discourse through dysfunctional engagement?

Urbano Reviglio (European University Institute)

Introduction

In the last years, 'doomscrolling' and 'brainrotting' have become mainstream terms referring, respectively, to the compulsive consumption of negative news and the deterioration of an individual's mental state due to excessive exposure to trivial content on social media. These are just two paradigmatic examples of broader risks associated with social platforms, which include ideological polarisation, the amplification of misinformation, and the erosion of democratic deliberation.

These risks arise from Very Large Online Platforms (VLOPs) (as defined in the EU Digital Services Act, DSA) and their ability to subtly shape user behaviour (whether through algorithmic covert influence or persuasive design) in order to optimise their engagement. That optimization logic, in turn, can undermine individual autonomy and degrade civic discourse. This article answers the following research questions: How the user engagement optimisation of VLOPs concur to systemic risks and how can VLOPs mitigate the risk that stem from this?

Context

For several years, concerns about the addictive nature of social media have intensified (Hari, 2022), yet the causal evidence remains limited and mixed. Tactics that capture user attention have been used since long in media (e.g., sensationalism), but today their persuasive power is far more granular and scalable. Disinformation spreads faster online, and platforms tailor capture strategies to individual users. I use "Dysfunctional Engagement" to denote engagement patterns that are compulsive, time-distorting and misaligned with users' goals.

Multiple forces contribute to this trajectory. Social media consumption surged during the COVID-19 pandemic, while news avoidance rose and plausibly harmed civic discourse. Short-form video became the dominant viral format, and the ensuing "TikTokification" of social media normalised such models. Public-health concerns about an "attention crisis" have grown, with documented correlates in minors such as insomnia and reduced sustained attention. More recently, the spread of "AI slop" and deepfakes has made engagement even more persuasive and harder to audit.

Despite expanding regulation (above all, the EU's DSA), users retain limited effective control over personalisation as well as time management. Articles 27 and 38 DSA are often cited as a basis for more user options and oversight, while Article 25 DSA (together with Article 5 AI Act) targets manip-

ulative designs (Reviglio & Fabbri, 2024). Yet no current legal regime requires platforms to demonstrate that popular tools (e.g., time limits) are fit for purpose and actually support user agency. If Article 25 DSA were interpreted more broadly, certain implementations could qualify as dark patterns, warranting stronger user controls (Reviglio & Giovanardi, 2024). This would align with the European Parliament's recent call to assess and restrict harmful, addictive techniques not covered by the Unfair Commercial Practices framework. Convincing evidence remains hard to obtain: researchers often rely on simulations and off-platform experiments that lack ecological validity (Leerssen, 2023).

Identifying credible design solutions therefore requires a better understanding of how recommender systems and interface choices causally produce Dysfunctional Engagement.

Methodology

Drawing on interdisciplinary evidence from HCI, behavioral science, and communication studies, and grounded in the normative aims and legal framework of the DSA, this article examines how VLOPs contribute to Dysfunctional Engagement, systematises open challenges, and evaluates potential solutions and research priorities. I conduct a comparative analysis of five social media VLOPs—TikTok, X, Instagram, Facebook, and YouTube—which are central to the European Commission's enforcement agenda. Methodologically, I combine: (i) a walkthrough analysis of design affordances (with a pre-specified coding scheme for capture, control, and transparency features); and (ii) a document review of the platforms' most recent audits and systemic-risk reports under the DSA.

Structure

The article proceeds as follows. First, I define Dysfunctional Engagement and map the systemic risks associated with recommender systems, detailing pathways from design choices to autonomy harms and civic degradation. Second, I analyse VLOP systemic-risk reports under the DSA to understand what they really are in this context (Loi et al., 2025), summarising identified risks and mitigation strategies while highlighting gaps and constraints in addressing Dysfunctional Engagement. Third, I examine the design affordances of TikTok, X, Instagram, Facebook, and YouTube to assess how specific features condition user control and choice architecture. Fourth, I synthesise the debate in functional tables that align: (a) risks and

mechanisms; (b) current evidence; (c) legal and design interventions; and (d) observed mitigation strategies across the selected VLOPs.

Objectives

This article analyses the interplay between risks to civic discourse and mental health in content consumption on VLOPs. By situating these dynamics within a systemic-risk framework (DSA Article 34), I map mechanisms through which platform design choices affect individual agency and the broader public sphere. This research contribute an interdisciplinary review of evidence on the externalities of engagement optimisation and a structured set of hypotheses and research directions for governance and design. It concludes by proposing a research agenda to quantify trade-offs between engagement, well-being, and democratic quality across different VLOP governance configurations.

Various research questions will be raised such as: What is the evidence that VLOPs' recommender systems and design concur to phenomena such as news avoidance, 'doomscrolling' and 'brainrotting'? To what extent certain options provided by VLOPs such as the one related to time management are effective? Should VLOPs apply a ratio of positive/negative content? Should they demote specific content? Which one? What design choices may qualify as dark patterns? And what pro-social nudges VLOPs could design to prevent such risks?

References

- Hari, J., *Stolen focus: Why you can't pay attention—and how to think deeply again*, Crown Publishing Group, New York, 2022.
- Leerssen, P., 'An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation' *Computer Law & Security Review*, Vol. 48, 2023, <https://doi.org/10.1016/j.clsr.2023.105790>.
- Loi, M., Fabbri, M. & Ferrario, A., 'Regulating the Undefined: Addressing Systemic Risks in the Digital Services Act (with an Appendix on the AI Act)', *Philosophy & Technology*, Vol. 38, 2025, <https://doi.org/10.1007/s13347-025-00903-7>.
- Reviglio, U., & Fabbri, M. Navigating the Digital Services Act: Scenarios of transparency and user control in VLOPs' recommender systems, *NORMALize 2024: The Second Workshop on the Normative Design and Evaluation of Recommender Systems*, 2024, <https://dx.doi.org/10.2139/ssrn.5040307>.

Reviglio, U., & Giovanardi, M., Advancing 'prosocial tech design' and shaping the EU's platform design governance, European University Institute, 2025, <https://doi.org/10.2870/3164689>.

Investigating the harms of viewing online misogyny on young people

Delali Konu (University of York), Carl Bunce (University of Reading), Eleanor Willcox (University of Oxford), Paul Galdas (University of York), David Zendle (University of York), Harriet Over (University of York)

Background

Very large online platforms including YouTube, Instagram, Tik Tok and X, offer many benefits to users but they can be used by extremists to spread damaging messages and amplify division within

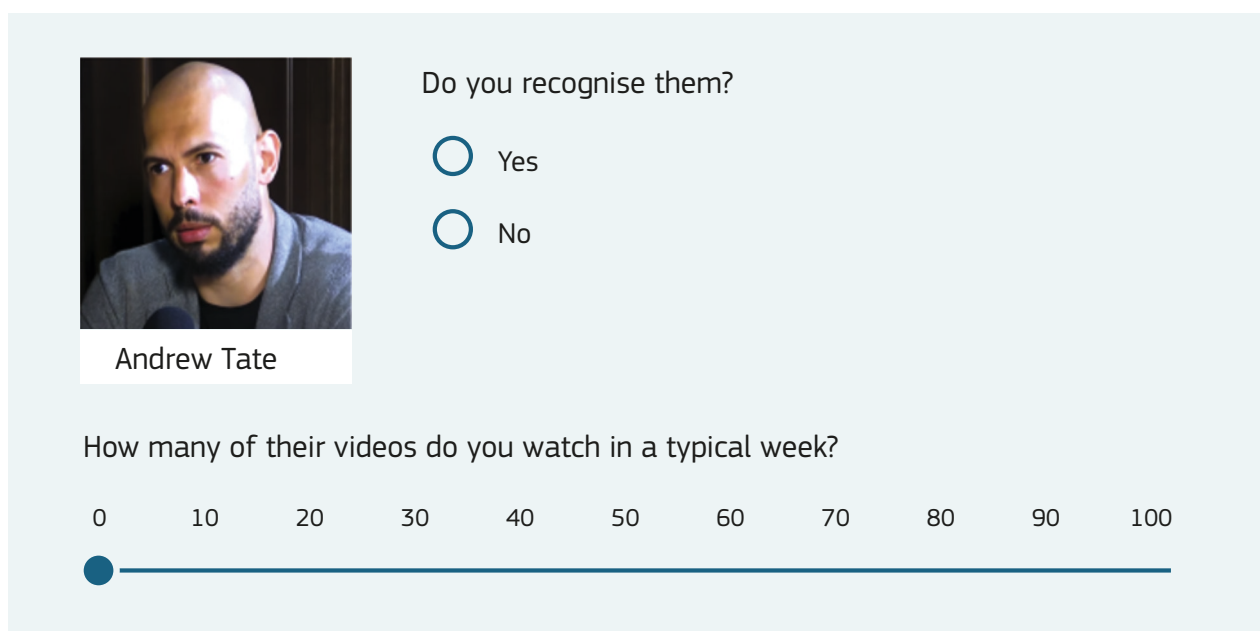
society (Ebner, 2023). This is evident in the manosphere, an online community that promotes male supremacy (Over, Bunce, Konu, et al., 2025).

The manosphere is made up of influencers from various subcommunities including, Men's Rights Activists, Men Going Their Own Way, Pick-Up-Artists, Involuntary Celibates and Manfluencers (Ging, 2019). These groups are united by their belief that men are innately superior to women and that men are the primary victims of injustice in contemporary society (Haslop & O'Rourke, 2021).

Their content is often distributed via monetised media including blogs, podcasts and courses (Bujalka et al., 2022). Algorithms have propelled this content into the mainstream with studies showing that, even without searching for this content, 69% of teenage boys aged 11 - 14 within the UK have been shown negative content about women and girls (Global Action Plan, 2024).

Parents, educators and policy makers are increasingly sounding the alarm about how the manosphere is affecting young people. Our previous research has shown that engagement with

FIGURE 3
Example item from the manosphere engagement task



The image shows a survey question interface. On the left is a portrait of Andrew Tate with his name 'Andrew Tate' written below it. To the right of the portrait is the question 'Do you recognise them?' followed by two radio button options: 'Yes' and 'No'. Below this is another question: 'How many of their videos do you watch in a typical week?' followed by a horizontal slider scale from 0 to 100 in increments of 10. The slider is currently set at 0.

Source: Authors' own elaboration

Image credit: Anything goes with Jamie English, 2023, via Anything Goes With James English - YouTube

the manosphere within schools is associated with increased discrimination towards female pupils and female teachers (Over, Bunce, Baggaley, et al., 2025).

Our new EU-funded project, HATESHIELD, seeks to 1) catalogue the ways in which the manosphere is affecting the behaviour and experiences of teenagers and young people 2) develop and evaluate research-led interventions to help young people think critically about online content.

Rationale for the current research

In the poster presentation, I plan to present the findings from two of our first studies in this area:

1) How young men's engagement with the manosphere affects its primary targets, women, by considering how viewing online misogyny relates to the treatment of women in romantic relationships.

2) How viewing online misogyny might affect its primary consumers, young men, in relation to their mental health.

These studies are a step towards understanding the impact of online misogyny on people's attitudes and behaviour. Both studies were given ethical approval by the departmental ethics committee, pre-registered and are due to be submitted for publication.

Methods

The first challenge in answering these questions is to reliably measure engagement with the manosphere. Polling data suggests that 80% of 16 to 17 year-old boys in the UK have watched content created by Andrew Tate, a prominent manosphere influencer (Hope not Hate, 2023) and 24% of young men agree with Andrew Tate's views about women (Smith, 2023). However, the manosphere is much broader than Andrew Tate.

We created a tool, the Manosphere Engagement Task, to measure individual engagement with influencers associated with the manosphere. Participants are asked to self-report their viewing behaviour of prominent manosphere influencers across social media platforms.

They view the names and pictures of 8 manosphere influencers and 4 social media influencers who are not associated with the manosphere (see Figure 3). They indicate how many of each influencer's posts they view weekly.

The task has high internal reliability ($\omega=.89$) and so allows us to measure the correlation between engagement with the manosphere and psychologically meaningful variables.

Results

In both studies, we analysed our data using a hurdle model. In our study on manosphere engagement and attitudes towards intimate partner relationships, we found that young men ($N=406$) who watch online misogyny are more likely to use coercive and controlling behaviours in their own relationships with women than young men who do not watch online misogyny ($z = 2.42, p=.016$). The more online misogyny young men watch, the more strongly they use coercive and controlling behaviours within intimate relationships ($z = 4.53, p<.001$).

In our study on manosphere engagement and mental health experiences we found that stigma against help seeking is related to whether young men ($N= 405$) watch manosphere content or not ($z = 3.41, p<.001$). Men who watch the manosphere express more stigma around seeking help for mental health problems.

Conclusions

Our results show that engagement with the manosphere is associated with negative attitudes towards women and girls within intimate partner relationships. It is thus possible that the growing popularity of the manosphere is one causal factor in increasing violence against women and girls.

Viewing online misogyny is associated with negative consequences for young men. In this particular study, men who engaged with the manosphere showed more problematic attitudes surrounding treatment and support for mental health problems. In the coming years, we will build on these findings with longitudinal research. We will also develop and refine new data donation techniques for more accurately measuring engagement with the manosphere. These study streams will help inform interventions to reduce the prevalence of problematic gender attitudes among teenagers and young people.

References

Bujalka, E., Rich, B., & Bender, S., 'The manosphere as an online protection racket: How the red pill monetizes male need for security in modern soci-

ety', *Fast Capitalism*, Vol. 19, Issue 1, 2022, <http://doi.org/10.32855/fcapital.202201.001>.

Ebner, J., *Going Mainstream: How extremists are taking over*, Bonnier Books UK, London, 2023.

Ging, D., 'Alphas, betas, and incels: Theorizing the masculinities of the manosphere', *Men and Masculinities*, Vol. 22, Issue 4, 2019, pp. 638–657, <https://psycnet.apa.org/doi/10.1177/1097184X17706401>.

Global Action Plan and Vodafone, 'AI "Aggro-rhythms": Young boys are served harmful content within 60 seconds of being online', 6 February 2024, <https://www.vodafone.co.uk/newscentre/press-release/ai-aggro-rhythms/>.

Haslop, C., & O'Rourke, F., (2021). "I mean, in my opinion, I have it the worst, because I am white. I am male. I am heterosexual": Questioning the inclusivity of reconfigured hegemonic masculinities in a UK student online culture', *Information, Communication & Society*, Vol. 24, Issue 8, pp. 1108-1122, <https://doi.org/10.1080/1369118X.2020.1792531>

Hope not Hate, 'Andrew Tate', 2023, <https://hope-nothate.org.uk/andrew-tate/>

Over, H., Bunce, C., Baggaley, J. & Zendle, D., 'Understanding the influence of online misogyny in schools from the perspective of teachers', *PLoS ONE*, Vol. 20, Issue 2, 2025, <https://doi.org/10.1371/journal.pone.0299339>.

Over, H., Bunce, C., Konu, D., & Zendle, D., 'Editorial Perspective: What do we need to know about the manosphere and young people's mental health?', *Child and Adolescent Mental Health*, 2025, <https://doi.org/10.1111/camh.12747>.

Smith, M., 'How many Britons agree with Andrew Tate's views on women', YouGov UK, 23 May 2023, <https://yougov.com/en-gb/articles/45735-how-many-britons-agree-andrew-tates-views-women>.

Studying minors and machines: rethinking longitudinal research on algorithmic influence

Alexandra Weilenmann (University of Gothenburg), Alan Said (University of Gothenburg)

Introduction: Systemic risks and the challenge of minors

The Digital Services Act (DSA) identifies systemic risks from very large online platforms and search engines that can affect public health, civic discourse, and individual rights. Article 34 requires designated platforms to assess and mitigate such risks – with the mental and physical health of minors explicitly recognised as a priority area.¹

Minors represent a high-risk group in digital environments: their cognitive, social, and emotional capacities are still developing, they are embedded in peer networks mediated by platforms, and they are disproportionately exposed to persuasive design.²

There is ongoing discussion among researchers, policymakers, and advocacy organisations about the impact of digital design features on minors' health and well-being. For example, the *Disrupted Childhood* report² describes elements such as variable rewards and infinite scrolling as common in children's digital environments and encourages further investigation of their developmental effects. Similarly, the *Better Internet for Kids* initiative defines "persuasive design" as techniques intended to increase user engagement, noting their relevance when considering children's time and activity online.³

Yet, they are also among the most difficult populations to study longitudinally. Legal consent requirements, ethical constraints, rapid changes in their online lives, and shifting platform affordances make it hard to collect robust evidence about long-term effects. In the case of AI use among minors, we are at a moment when adoption is still emerg-

NOTE

¹ <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act>

² <https://rightsfoundation.com/resource/updated-report-disrupted-childhood-the-cost-of-persuasive-design/>

³ <https://better-internet-for-kids.europa.eu/en/learning-corners/teachers-and-educators/persuasive-design>

ing, offering a rare opportunity to design longitudinal research coherently from the outset.

Insights from longitudinal and cross-platform studies

We draw on our empirical studies, alongside related work, to illustrate how young people's interactions with AI chatbots, social media, and recommendations evolve over time.

Longitudinal trust in Snapchat's My AI

In a four-week qualitative study with 27 young adults (under review), we examined how trust in Snapchat's My AI (a GPT-powered chatbot) evolved over repeated interactions. We identified two sets of influences:

- Chatbot-related factors: ability/expertise, conversational style, human-likeness.
- Environment-related factors: transparency/privacy, perceived risk, platform context.

Although participants valued the immediacy and friendliness of the chatbot, some also expressed uncertainty about data use transparency and how personalisation was achieved (Vanhoffelen, 2025). Changes in trust over time appeared to be influenced by both the chatbot's behaviour and the broader platform context, including pre-existing attitudes toward the service.

For minors, these patterns may be more pronounced, as younger users might attribute more human-like qualities to chatbots, have different perceptions of privacy risks, and be more attuned to platform norms ⁴.

A decade of social media use

Our recent study followed the same group of social media users over ten years, documenting shifts in usage patterns from public to private, active to passive, and enjoyment to more complex or problematic experiences (Jungselius & Weilenmann, 2025). The study highlights that shifts in design, e.g., algorithmically curated feeds, and ephemeral

formats, can shape user practices alongside individual preferences.

Other research also links changes in social media use to mental health trajectories in adolescence. For example, a longitudinal cohort study found that within individuals, increases in social media use were associated with increases in depressive symptoms across early adolescence (Nagata et al., 2025).

While this does not establish causality, it illustrates the importance of tracking both behavioural changes and broader well-being outcomes over time. UNICEF further notes that the quality, context, and purpose of digital use may be more meaningful indicators than overall duration alone ⁵.

For minors, these structural shifts may influence the developmental trajectory of social media use in important ways. For example, the move from public posting to smaller-group sharing may reduce some exposure to wider audiences but could also create more closed peer spaces. Similarly, a shift from active creation to more passive consumption may alter perceptions of agency and increase reliance on algorithmically selected content (Jungselius, 2024).

Autonomy, recommender systems, and the intention-behaviour gap

Our work on prediction accuracy and autonomy in recommender systems (Angwald et al., 2021) highlights the intention-behaviour gap: what users do on a platform (implicit data) often diverges from what they intend to do (explicit preferences). For minors, this gap may be even wider due to developmental factors affecting self-regulation and evolving media habits (Orben & Blakemore, 2023).

Systems optimised for engagement may, in some cases, amplify certain patterns of use that are not aligned with user intentions. Understanding how such systems interact with design features and usage practices requires careful, multi-method study designs capable of distinguishing between different types of engagement and their outcomes.

NOTE

⁴ <https://saferinternet.org.uk/blog/how-does-snapchats-new-ai-function-my-ai-impact-young-people>

⁵ <https://www.unicef.org/innocenti/reports/childhood-digital-world>

Why correlation is not causation – especially with minors

A recurring challenge in this research area is the tendency to over-interpret associations between observed behaviours and specific outcomes. Apparent correlations in behaviour should not be mistaken for causation. For instance, high engagement with a chatbot may reflect novelty or lack of alternatives rather than trust; similarly, shifts toward passive consumption or extended watch time may stem from algorithmic design choices rather than user preference.

For minors, careful interpretation is particularly important, as studies suggest that the ability to critically evaluate persuasive elements in media continues to mature through adolescence (Orben et al., 2024). This implies that design choices and algorithmic curation may have effects on attitudes, habits, and self-concept that are not immediately visible and may require sustained observation to understand fully.

Rethinking longitudinal studies with minors

Designing longitudinal research involving minors requires approaches that not only address ethical and legal constraints but also capture the complexity of their interactions with evolving digital technologies. A practical starting point is mixed-method triangulation: combine qualitative methods (e.g., diaries, interviews, stimulated recall) and quantitative approaches with access to platform data (e.g., under DSA research provisions) so that studies can record both what happens and how it is experienced over time. This helps connect fine-grained behaviour logs with situated accounts of meaning-making in everyday life.

Because digital environments are anything but static, studies should explicitly account for platform affordances and design features. Track affordances such as visibility, persistence, and algorithmic curation, as well as interface patterns that may change during the study; doing so helps distinguish behavioural shifts driven by design changes from those emerging from participants' own development (Angwald et al., 2021; Kaun & Stiernstedt, 2014).

Closely related is the need to clarify what is being studied: following the “screen-time mayhem” critique, avoid collapsing diverse activities and con-

texts into a single measure; specify activity types, contexts, and purposes so findings remain interpretable and cumulative (Kaye et al., 2020).

Equally important are ethical and practical safeguards tailored to minors. When involving young people in shaping research questions or interpreting findings, structure participation to encourage critical reflection and ensure diverse perspectives, reducing the risk that prevailing societal narratives (e.g., that smartphones or social media are inherently harmful) dominate the process.

Maintain transparent research practices consistent with standards for work with children and incorporate adaptive consent so assent/consent evolves as minors mature and their understanding of risks changes over time (Helgesson, 2005).

Finally, use transparency mechanisms (e.g., the DSA transparency database, ad repositories, and research data access channels) to complement self-reports and opaque aggregates, strengthening the evidentiary basis without over-reliance on any single data source. Designing longitudinal research involving minors requires approaches that not only address ethical and legal constraints but also capture the complexity of their interactions with evolving digital technologies.

Conclusions: towards evidence that matters for minors

Understanding systemic risks to minors in algorithmic environments requires approaches that balance methodological rigor, ethical considerations, and sensitivity to developmental contexts. Longitudinal designs remain essential but must adapt to the realities of studying minors: legal constraints, rapidly changing platforms, and evolving user behaviour. Understanding systemic risks to minors in algorithmic environments requires approaches that balance methodological rigor, ethical considerations, and developmental sensitivity.

We propose a research agenda that treats correlations as prompts for deeper inquiry rather than proof of causation, and that combines transparency data with qualitative, developmental, and participatory insights. Such an approach can support the DSA's systemic risk objectives while generating insights that meaningfully address minors' experiences in digital environments. In the case of AI use among minors, we are at a moment when adoption is still emerging, offering a rare oppor-

tunity to design longitudinal research coherently from the outset.

By learning from the methodological challenges that have limited past “screen time” research – such as over-reliance on self-reports, simplistic time-based measures, and a lack of clarity on what is actually being studied (Kaye et al., 2020) – researchers can focus on mapping platform affordances (Bucher & Hellmond, 2018), design features, and usage contexts that shape AI interactions in practice (Orben et al., 2024). This will allow for a more precise understanding of how technology use unfolds over time, avoiding the pitfalls of earlier approaches and building an evidence base that is both robust and relevant to policy and practice.

References

Angwald A., Areskoug K. & Said A., ‘Prediction accuracy and autonomy’, Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop 2021, 2021, <https://doi.org/10.48550/arXiv.2211.08134>.

Bucher T. & Hellmond A. ‘The affordances of social media platforms’, in: Burgess J., Marwick A. & Poell T. (eds.), *The SAGE Handbook of Social Media*, SAGE Publications, London, 2018.

Helgesson G. ‘Children, longitudinal studies, and informed consent’, *Medicine, Health Care and Philosophy*, Vol. 8, Issue 3, pp. 307–313, 2005, <https://doi.org/10.1007/s11019-005-0978-4>.

Jungselius B., ‘A scoping review of current research on social media use among children and adolescents’, *Discover Psychology*, Vol. 4, Issue 1, 2024 <https://doi.org/10.1007/s44202-024-00226-2>.

Jungselius B. & Weilenmann A., ‘Tracing change in social media use: A qualitative longitudinal study’, Proceedings of the CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–14, <https://doi.org/10.1145/3706598.3713813>.

Kaye L., Orben A., Ellis D., Hunter S. & Houghton S., ‘The conceptual and methodological mayhem of “screen time”’, *International Journal of Environmental Research and Public Health*, Vol. 17, Issue 10, 2020, <https://doi.org/10.3390/ijerph17103661>.

Kaun A. & Stiernstedt F., ‘Facebook time: Technological and institutional affordances for media memories’, *New Media & Society*, Vol. 16, Issue 7, 2014, <https://doi.org/10.1177/1461444814544001>.

Nagata, J.M., Otmar, C.D., Shim, J., Balasubramanian, P., Cheng, C.M., et al., ‘Social Media Use and Depressive Symptoms During Early Adolescence’, *JAMA Network Open*, Vol. 8, Issue 5, 2025, <https://doi.org/10.1001/jamanetworkopen.2025.11704>.

Orben, A. and Blakemore, S.-J., ‘How social media affects teen mental health: a missing link’, *Nature*, Vol. 614, 2023, pp. 410–412. <https://doi.org/10.1038/d41586-023-00402-9>.

Orben, A., Meier, A., Dalgleish, T. and Blakemore, S.-J., ‘Mechanisms linking social media use to adolescent mental health vulnerability’, *Nature Reviews Psychology*, Vol. 3, Issue 6, 2024, <https://doi.org/10.1038/s44159-024-00307-y>.

Vanhoffelen, G., Vandenbosch, L. and Schreurs, L., ‘Teens, Tech, and Talk: Adolescents’ Use of and Emotional Reactions to Snapchat’s My AI Chatbot’, *Behavioral Sciences*, Vol. 15, 2025, <https://doi.org/10.3390/bs15081037>.

Systemic risks to minors on online platforms: evidence from three studies by Landesanstalt für Medien NRW

Maïke Resing (Landesanstalt für Medien NRW), Meike Isenberg (Landesanstalt für Medien NRW)

Introduction

The Landesanstalt für Medien NRW is one of 14 independent German media authorities. It is based in the federal state of North Rhine-Westphalia (NRW), which is the most populous federal state in Germany. Our responsibilities include the oversight of private radio broadcasters, television broadcasters, and electronic media. This encompasses issuing broadcasting licences and supervising compliance with the provisions set forth in the German Interstate Broadcasting Agreement, the Interstate Treaty on the Protection of Minors from Harmful Media, and the state media laws and treaties, or the European AVMS Directive, respectively.

As online platforms have assumed an ever more prominent role in media consumption and public communication, their regulation and the study of their impact on users, particularly minors, have

become a key focus of our work. Online platforms have become integral to the daily lives of children and adolescents. While offering opportunities for communication and creativity, these platforms also expose minors to significant risks. The European Centre for Algorithmic Transparency (ECAT) has emphasised the need to investigate systemic risks posed by very large online platforms (VLOPs) and very large online search engines (VLOSEs), especially those affecting young users. This paper presents findings from three recent studies conducted by Landesanstalt für Medien NRW that highlight systemic risks to minors' mental and physical health on VLOPs.

Study 1: Cyber grooming on social media (2025)

Conducted annually since 2021, the representative online survey of more than 2,000 children and adolescents aged 8 to 17 provides valuable insights into the evolving dynamics of cyber grooming. The latest survey shows that nearly one in four minors has experienced cyber grooming. The most common tactics include offenders attempting to arrange in-person meetings with minors or persuading them to send sexually explicit images or videos – often by offering rewards or incentives in exchange. The platforms most frequently used by offenders to contact minors were Snapchat (29%) and Instagram (27%), highlighting the need for stronger safeguards on these services.

After gaining the trust of young people, offenders use collected material (e.g., chat logs, images, or videos) to blackmail victims into further actions. According to the study, 26% of those affected by cyber grooming had already experienced sextortion. The study also found that many children, especially younger ones, do not disclose these experiences to anyone. These findings underscore the role of platforms in facilitating grooming and the urgent need for protective mechanisms.

Study 2: Porn consumption and sexting (2024)

The representative study on pornography and sexting among youth shows that 42% of more than 2,800 11-17-year-olds had encountered pornographic content online. That is 7% more compared to our previous survey in 2023. Alarmingly, this rise is concentrated among children aged 11 to 13, where the proportion increased from 19% in 2023 to 26% in 2024. Many minors encounter

pornography typically between the ages of 12 and 15 for the first time. Often, this early exposure occurs unintentionally.

This trend highlights growing accessibility and exposure at increasingly younger ages. A substantial number of respondents expressed discomfort and confusion regarding the content they had seen. The lack of contextual understanding and absence of guidance contribute to emotional distress and potential long-term effects on sexual development. This raises serious concerns about the accessibility of explicit content and the lack of effective age-verification mechanisms online.

Study 3: TikTok challenges (2024)

Combining a content analysis of over 2,500 TikTok videos with a quantitative survey of 755 10-16-year-olds, this study examined the prevalence and risks of viral challenges. The content analysis showed that the majority – around 65% – of the challenge videos examined were harmless TikToks, such as dance or singing videos. However, around one third of the videos show potentially harmful challenges, and 1% even show potentially deadly challenges.

The analysis shows that videos with negative content do not necessarily achieve a higher reach than other videos. TikTok regulates harmful content by blocking both the challenges and related search terms. However, if a harmful video is not regulated early enough or sufficiently by TikTok, the platform's algorithm allows challenges to spread rapidly within a few weeks. The study criticises TikTok's inadequate and opaque moderation practices, which allow harmful content to proliferate. Algorithmic amplification of such content exacerbates the danger, making this a clear case for platform auditing under the DSA.

Relevance to the Digital Services Act (DSA)

Our research highlights the prevalence of cyber grooming, exposure to pornography at increasingly younger ages, and the spread of dangerous challenges on social media platforms. This highlights the urgent need for greater transparency in how very large online platforms operate, particularly in relation to minors' exposure to harmful content.

Despite the regulatory framework established by the DSA, platforms continue to fall short in providing sufficient insight into their content moderation

practices, algorithmic amplification mechanisms, and age-verification systems. Article 40 of the DSA is therefore not merely a procedural provision, but a critical enabler of accountability. Platforms must facilitate meaningful data access for vetted researchers, allowing them to audit risk factors and evaluate the effectiveness of platform safeguards.

Transparency in practice? Platform moderation trends after Romania's annulled presidential election

Viktor Kaupp (Maastricht University), Carl-Anton Lüninck (Maastricht University), Elizabeth Mataj (Maastricht University), Henry Tari (Maastricht University), Catalina Goanta (Utrecht University), Adriana Iamnitchi (Maastricht University)

Introduction

In December 2024, Romania annulled its presidential election following widespread concerns over foreign interference, with particular attention to the role of TikTok in disseminating political content. Soon after, the European Commission opened a formal investigation under the Digital Services Act

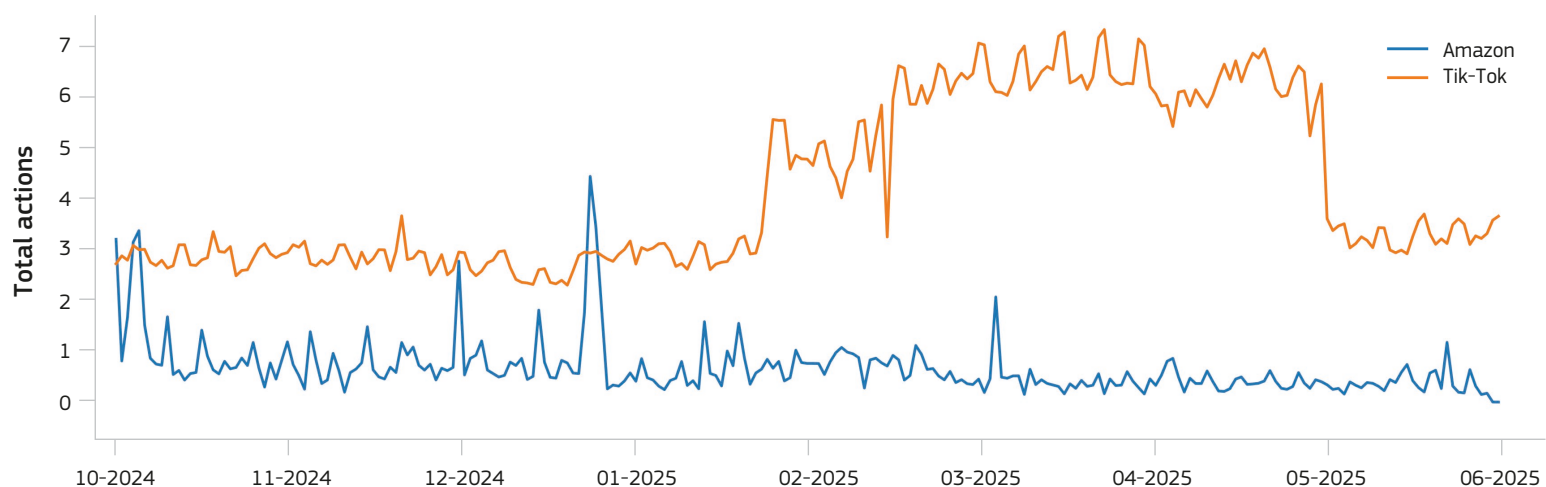
(DSA), citing potential violations related to content moderation, transparency, and electoral integrity [Commission, 2024]. Since September 2023, the very large online platforms (VLOPs) publish daily reports of their moderation actions in the publicly accessible DSA Transparency Database [European Commission, 2023].

Although this database provides a standardized stream of information across platforms, early analyses [Kaushal et al., 2024, Trujillo et al., 2025] identified serious limitations that hinder its utility for transparency. In particular, key metadata fields—such as the language of the moderated content or the specific violation category—are often missing, inconsistently reported, or deliberately vague (e.g., the frequent use of “Terms of Service” as a generic violation label).

This study investigates whether, two years after the database’s launch and in the aftermath of the European Commission’s investigation into systemic risks, very large social media platforms have modified their reporting practices. We critically assess the suitability of the DSA Transparency Database as a tool for detecting and analyzing platform-based interference in democratic processes.

Taking the annulled 2024 Romanian presidential election—the first European election overturned due to social media interference—as a stress test, we examine whether moderation data from TikTok,

FIGURE 4
Daily moderation volume, TikTok vs. Amazon



Source: Authors' own elaboration, based on data from the DSA Transparency Database

Meta, and X provide actionable insights into platform behavior around major democratic events.

While our analysis reveals broad shifts in moderation practices—for example, Meta’s move from removals to demotions and TikTok’s sudden reassignment of violation categories—the database ultimately fails to support event-specific or country-level analysis. Its most significant limitations continue to include the absence of geographic and linguistic metadata, strategic ambiguity in categorization, and the lack of independent verification due to reliance on self-reported data.

To address these limitations, we also adopt an alternative methodology: rather than focusing on moderation data, we analyze changes in the overall volume and timing of moderation records. Our hypothesis is that even without knowing precisely what was moderated, patterns in how much and how quickly moderation occurred may still reveal responsiveness to regulatory or political pressures.

To test this, we apply both machine learning techniques for changepoint detection and statistical causal inference (Difference-in-Differences) to evaluate whether real-world events significantly influenced TikTok’s moderation behavior, using Amazon Shopping — a platform with minimal political exposure — as a control case.

Methodology

This study relied on data from the DSA Transparency Database, which records Statements of Reasons (SoRs) reported by major platforms. We extracted datasets from TikTok, Facebook, Instagram, and X covering the period 1 October 2024 – 31 May 2025, which included both the annulled Romanian presidential election (December 2024) and the repeat election (May 2025). In addition, we used the statements of reasons reported by Amazon Store (as the control signal), an e-commerce platform with no known political content exposure, serving as a baseline for non-election-related moderation behavior.

We initially explored several other large online platforms as potential control groups, including Uber and Booking.com. However, their moderation patterns revealed structural differences: e.g., they exhibited no moderation activity on weekends, suggesting manual-only moderation workflows. Such patterns would introduce systematic bias into Difference-in-Differences (DiD) analysis, as changes might reflect staffing schedules rather than exogenous events. Amazon Store was ulti-

mately selected because its moderation appeared continuous and unaffected by political events. Figure 4 shows the raw daily moderation volumes for TikTok and Amazon over the study window.

Two periods were defined for comparison: Period I: 15 October – 15 December 2024 (pre-annulment); and Period II: 4 January – 31 May 2025 (post-annulment). We downloaded over 80GB of raw data and processed it using Python. We computed moderation delay as the difference between content creation date and moderation date. Key attributes we analyzed included moderation delay, automation (detection vs. decision), content type, decision ground (illegal vs. incompatible content), decision visibility (removed, demoted, disabled, labeled), and violation category.

Categories relevant to elections were prioritized: Negative effects on civic discourse or elections, Illegal or harmful speech, and Risk for public security. A major challenge was the absence of usable content language and territorial scope metadata. Attempts to approximate Romanian-specific moderation using text fields and fast-Text classification yielded no Romanian content, highlighting a critical gap in the transparency framework.

To analyse changes in the moderation volume that would signal change in moderation practices, we developed a memory-aware data collection and preprocessing approach. Daily compressed archives were downloaded from the DSA Amazon S3 bucket using a custom Python pipeline. Each archive was recursively unpacked, CSVs loaded in low-memory mode, and relevant fields standardized. Two key dates were extracted: content creation and moderation application. From these, we computed daily metrics: (i) the total number of moderation actions, and (ii) the delay in days between content creation and moderation.

Aggregation by day and platform was performed, trimming the first and last five days to avoid distortions. While mean moderation delay was initially considered as an indicator of responsiveness, it proved unreliable due to archival moderation of old content; the analysis therefore focused on total actions. To identify structural breaks in moderation behaviour, we applied the Pruned Exact Linear Time (PELT) algorithm via the ruptures library. The data were pre-processed with a 7-day rolling average and Z-score normalization. We explored multiple penalty levels (β {10, 20, 30, 40}), selecting changepoints [Truong et al., 2020] that appeared consistently across higher penalties as robust indicators.

To estimate causal effects, we employed a Difference-in-Differences (DiD) framework [Angrist and Pischke, 2009] with TikTok as the treated unit and Amazon Shop as control. The model regressed Z-scored daily moderation counts on event timing, treatment status, and their interaction, with the interaction term capturing TikTok's relative shift.

Empirical observations

The lack of localization (country and language metadata) made it impossible to identify Romanian-specific content. While platforms exhibited broad shifts in moderation, no direct connection to the Romanian elections could be established. TikTok reported over 1.01 billion moderations, primarily for illegal or harmful speech and scope of platform services. Election-related removals (43 million) were not linked to Romania. Between Period I and II, moderation volume increased by 48%, with consistently high automation (99%).

Moderation delays shortened, suggesting efficiency improvements. Despite public claims of removing 27,000 Romanian interference accounts, no such actions appeared in the DSA database, implying reclassification under vague categories. Facebook reported 438 million moderations, but only 528 cases were labeled as election-related. The majority (>90%) fell under the ambiguous scope of platform services. A clear strategic pivot occurred: in Period I, 93% of the moderated records were removals. In Period II, 81% of the statement of reasons were demotions ("shadow banning"). This shift coincided with Meta's January 2025 announcement scaling back fact-checking.

In April 2025, category differentiation improved, with cases redistributed from scope of platform services to scams and fraud, harmful speech, and others. Moderation delays increased in Period II. Instagram reported 94 million moderations, with only 54 linked to elections. Overall moderation rose modestly (+7.8%). Unlike Facebook, Instagram maintained a stable distribution of removals (92%). Notably, manual moderation increased, with "not automated" decisions reaching 15.7%, the highest across platforms.

Moderation delays showed minor improvements. X reported extremely low moderation (3,000/day, all manual). Only one case was flagged for election interference. A three-week reporting gap occurred between March 25 and April 15, 2025. All actions were reported with zero delay, indicating possible misreporting. Overall, X's moderation appeared minimal and unreliable.

When comparing the timeseries of TikTok records with those submitted by Amazon Store, two robust breaks in TikTok's moderation activity were identified: 24 January 2025 and 4 May 2025. The January break aligns with the opening of European Commission proceedings against TikTok under the DSA. The January event produced a large, statistically significant increase in TikTok's moderation relative to Amazon (coefficient 2.254, $p < 10^{-40}$).

The May event, by contrast, was not significant (coefficient -0.014 , $p = 0.96$). The January surge suggests a platform-level response to regulatory pressure, even if causality cannot be proven. The May decline appears to reflect a transient fluctuation not specific to TikTok. These results demonstrate that macro-level signals of responsiveness can be detected, even without content-level or country-specific metadata.

The DSA database lacks reliable fields for language, geolocation, or topic, preventing Romania-specific analysis. Delay measures are distorted by moderation of archival content, while platform heterogeneity complicates control selection. Temporal correlation with political events suggests, but does not confirm, causal effects.

References

- Angrist J.D. & Pischke J.S., *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton NJ, 2009.
- European Commission, 'Commission opens formal proceedings against TikTok on election risks under the Digital Services Act', Press release, 17 December 2024, accessed 7 August 2025, https://ec.europa.eu/commission/presscorner/detail/en/ip_24_6487.
- European Commission, 'DSA Transparency Database', European Commission website, accessed 7 August 2025, <https://transparency.dsa.ec.europa.eu/>.
- Kaushal R., Van De Kerkhof J., Goanta C., Spanakis G. & Iamnitich A., 'Automated transparency: A legal and empirical analysis of the Digital Services Act transparency database', *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 1121–1132, <https://doi.org/10.1145/3630106.3658960>.
- Trujillo, A., Fagni, T., & Cresci, S., 'The DSA Transparency Database: Auditing self-reported moderation actions by social media', *Proceedings of the ACM on Human-Computer Interaction*, Vol.9, Issue 2, 2025, <https://doi.org/10.1145/3711085>.

Truong C., Oudre L. & Vayatis N., 'Selective review of offline change point detection methods', *Signal Processing*, Vol. 167, 2020, <https://doi.org/10.1016/j.sigpro.2019.107299>.

Waking up to smartphones: Towards evidence based solutions

Lisa Henderson (University of York), Emma Sullivan (University of York)

Background

Smartphones and social media have become an integral part of daily life, with nearly 5 billion users worldwide (Kemp, 2025). Whilst social media can provide opportunities for education, connection and entertainment, concerns persist regarding its impact on sleep, mental health and academic performance - particularly among young people. Research increasingly links problematic use to sleep disturbances, offering a potential causal mechanism for subsequent effects on mental health and academic performance; however, findings remain inconsistent due to methodological heterogeneity, reliance on self-report measurement, and an overemphasis on cross-sectional designs. In addition, growing debate surrounds the effectiveness of school smartphone bans, with limited evidence on whether such policies can improve sleep, wellbeing, or safeguarding outcomes.

To address these critical gaps, we will present research drawing on varying methodologies:

- (1) a scoping review of longitudinal research on social media and sleep,
- (2) feasibility and pilot trials of the effects of smartphone/social media detoxes on young people,
- (3) a national analysis of school smartphone policies and how they relate to safeguarding incidents in schools.

Methods

Scoping review: We synthesised longitudinal studies (2018-2023) examining the impact of social media use on downstream sleep in children and adolescents. Feasibility and pilot trials: We have implemented a 21-day social media and smart-

phone total detox intervention in UK secondary schools in two studies (the first of which was televised on C4 Documentary "Swiped: The School that Banned Smartphones", with over 1.5 million viewers in the two weeks following release).

Pre/post measures of sleep, wellbeing, and cognition were collected via surveys, wearable devices and cognitive tasks. Policy analysis: Freedom of Information (FOI) responses were gathered from 114 schools in England (>90,000 pupils) to examine the prevalence and content of smartphone policies during 2023/24 and their association with reported digital safeguarding incidents.

Results

Scoping review: Most longitudinal studies suggest that problematic or excessive social media use - particularly near bedtime - negatively influences sleep outcomes (e.g., delayed bedtime, increased sleep onset latency). However, reliance on self-report and non-validated measures limits causal inference.

Feasibility and pilot trials: Following the 21-day ban, pupils reported reduced sleep onset latency (~20 minutes) and increased total sleep duration (~1 hour). Wearable data mirrored these improvements. Gains were observed in attention but not working memory. Anxiety and negative mood decreased significantly, while depression symptoms showed small, nonsignificant reductions. Social connectedness was unchanged. These findings demonstrate the feasibility and potential benefits of collective social media and smartphone detoxes for sleep and wellbeing.

Policy analysis: Only 53% (60/114) of schools in England reported having a phone policy, with significant variability in restrictiveness and inconsistent use of "ban" rhetoric. Digital safeguarding incidents represented only 4.9% of total incidents (4183/85,540), but with substantial variation across schools (>10% of total incidents in 36/114 schools and >25% in 9), suggesting inconsistent reporting. Schools with phone policies (particularly more restrictive ones) reported more incidents, though most incidents occurred outside school. This suggests policy presence may increase detection and reporting rather than reduce incidents, and that school bans are not a total solution.

Conclusions & implications

Excessive and nighttime social media use is associated with poorer sleep, but stronger causal evidence is needed. Detox trials are one route to such causal evidence, and could also offer collective approaches to better digital education and safety and empower children and parents to make informed decisions. School-only bans appear feasible and may be linked to improved reporting of digital safeguarding incidents, but current UK school policy practice is highly inconsistent, with unclear effectiveness in reducing digital harms.

Greater standardisation and alignment across education, health, and home settings are required. A coordinated, multi-sector approach—co-designed with young people—is essential for promoting healthier digital behaviours and safeguarding youth wellbeing.

References

Kemp S., 'Digital 2025: Global overview report', DataReportal, 5 February 2025, <https://datareportal.com/reports/digital-2025-global-overview-report>.

X's Community Notes: Algorithmic resolution of crowd-sourced moderation on X in polarised settings across countries

**Paul Bouchaud (CNRS, EHESS, SciencesPo),
Pedro Ramaciotti (CNRS, SciencesPo, LPI)**

Social media platforms increasingly face the challenge of misinformation while balancing freedom of expression with content moderation. Traditional expert factchecking, while effective at reducing agreement with misleading claims, suffers from significant scalability limitations, potential partisan bias perceptions, and disconnection from on-line attention patterns.

These constraints have prompted platforms to explore crowd-sourced moderation systems, where regular users collaboratively identify and contextualise potentially misleading content. X pioneered this approach through its Community Notes system, initially piloted in the United States in 2021 and deployed globally in 2023. The system enables selfselected contributors to attach contextual notes to posts they consider misinformed or potentially misleading, with other contributors rating

these notes based on subjective assessments of content and context. To address the core challenge of moderating notes on polarised topics where consensus may not emerge across ideological divides, X employs a bridging-based algorithmic approach that learns an latent ideology for both users and notes, then identifies notes that appeal to users across ideological spectra rather than only to like-minded individuals.

X's Community Notes were shown, during the pilot phase in the United States, to garner broad ideological appeal and to reduce the sharing by users of flagged content. Yet, the system's performance within other political landscapes, in particular in European multi-party systems, remained unexplored.

This study examines Community Notes usage and outcomes across 13 countries, analysing 1.9 million moderation notes with 135 million ratings from 1.2 million users. We cross-reference the latent ideological space inferred by X's algorithm with established political dimensions derived from follower networks of Members of Parliament and calibrated with Global Party Survey data. The analysis encompasses geographically diverse countries including the United States, United Kingdom, Japan, Spain, France, Brazil, Canada, Germany, Argentina, Israel, Australia, Poland, and Mexico.

Our findings reveal that Community Notes usage exhibits highly unequal distributions globally, with the majority of notes relating to the United States, followed by Japan, the United Kingdom, Brazil, and France; see Figure 5. Political content accounts for 65.2% of posts and 76.6% of Community Notes ratings, with the most frequently discussed topics being Political Authority, Law and Order, Equality, Environmental Protection, and Freedom and Human Rights.

Furthermore, we observe that community notes are being requested, proposed and approved as "helpful" across the left-right political leaning in all examined countries, with country-specific imbalances displayed in Figure 6. Established media outlets constitute the predominant sources in notes, appearing in 25.9% of cases, followed by X posts (17.4%) and Wikipedia articles (8.9%), while expert fact-checking articles appear in only 3.5% of proposed notes. Yet notes referencing expert fact-checking are more frequently deemed "helpful" than those that do not.

The outcomes of the algorithmic resolution demonstrate significant limitations in the system's

effectiveness. Only 11.97% of proposed Community Notes arguing that posts are misinformed or potentially misleading garner sufficient favorable ratings from diverse contributors to attain “Helpful Status” and become visible on X. The remaining notes either achieve “Not Helpful Status” (2.81%) or lack sufficient ratings from diverse contributors (85.22%). Politically divisive accounts show substantially lower rates of notes achieving Helpful Status, with accounts like Elon Musk or Kamala Harris’s campaign having 3% or less than 0.1% of their annotated posts receiving such notes, respectively.

Crucially, notes discussing politically divisive issues consistently struggle to garner consensus across all analysed countries. In the United States, only 6.0% of notes discussing presidential candidates and political parties, 5.5% addressing the Israel-Palestine conflict, and 11.7% covering the Russia-Ukraine war reach Helpful Status, compared to 39.0% of notes reporting scams, fraud, and impersonation. This pattern reflects the system’s fundamental design limitation: its reliance on cross-partisan consensus inherently prevents resolution of the most polarising content.

The latent ideological dimension learned by X’s Community Notes system demonstrates varying degrees of alignment with established political dimensions across countries. In the United States, the single-dimensional latent space closely aligns with Left-Right political leanings. Across all 13 countries, user positions along the primary political dimension predict the sign of latent ideology θ_n with AUC values ranging from 0.850 in Poland to 0.729 in Israel, averaging 0.808 ± 0.037 .

These findings provide the first empirical validation of X’s Community Notes design hypothesis in a global comparative context while simultaneously revealing critical limitations. The system effectively captures each country’s main polarising dimension but fails, by design, to moderate the most polarising content, creating potential risks to civic discourse and electoral processes.

X’s Community Notes emerge from our results as falling short of their mission during events that activate structural political divisions, such as national elections, leading to potentially negative effects on civic discourse, electoral processes, and public security.

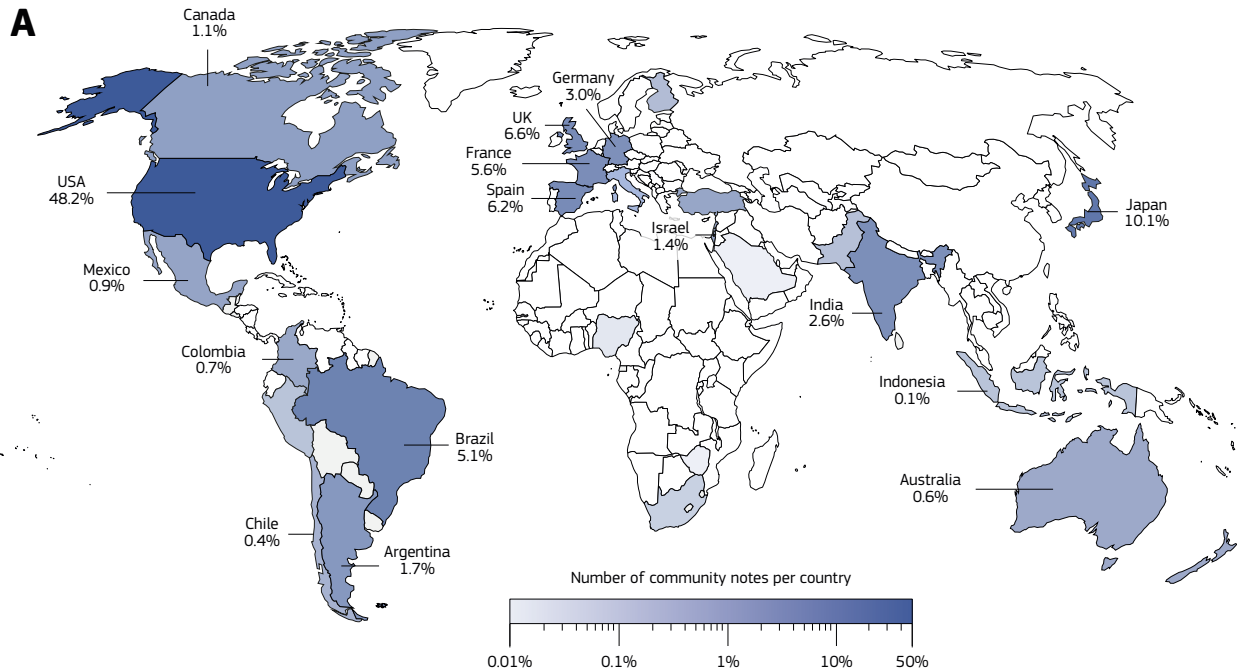
As such, the inherent limitations in consensus-based approaches to moderating misinformation should be acknowledged in platform risk

assessments conducted under Article 34.1.c of the Digital Services Act, particularly as platforms like YouTube, TikTok, and Meta deploy similar crowd-sourced moderation systems. To mitigate these potential negative effects on civic discourse, complementing crowd-sourced approaches with expert fact-checking, as originally intended by the original Twitter teams, may appear appropriate.

FIGURE 5

Community Notes usage and outcomes:

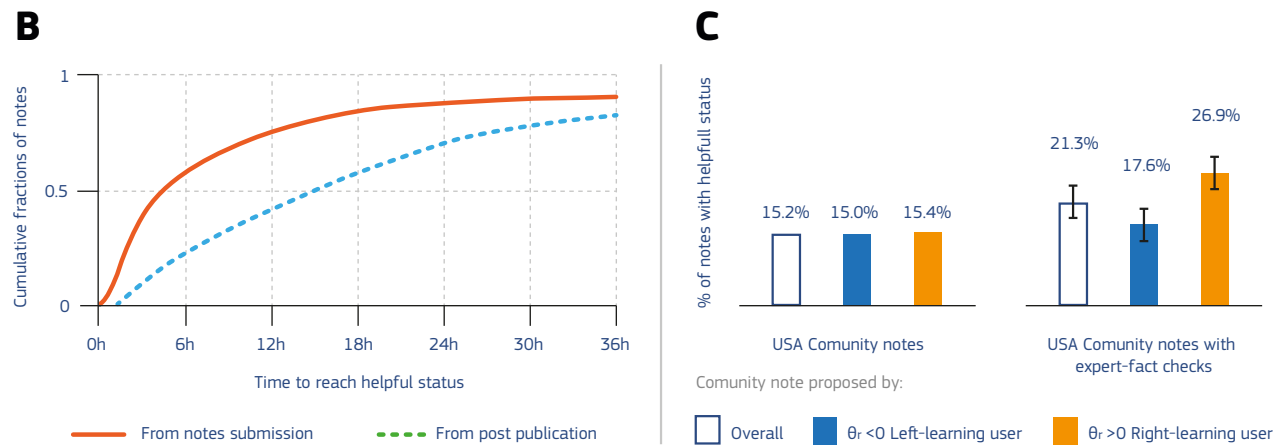
(A) Fraction of Community Notes proposed per country.



Community Notes usage and outcomes:

(B) Time distribution for notes to reach Helpful Status.

(C) Fraction of notes reaching Helpful Status in the US by ideological leaning of authors, with and without expert fact-checking references.



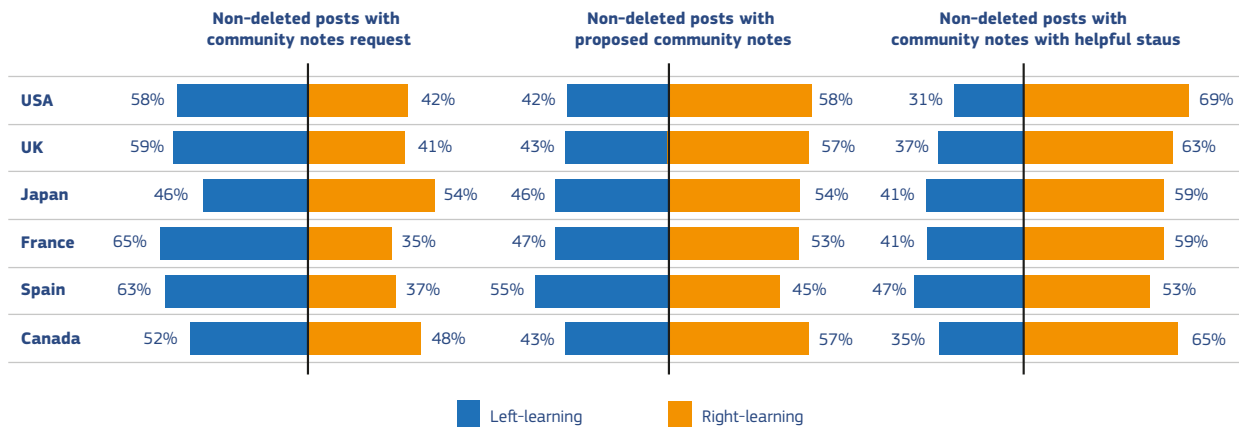
Source: Authors' own elaboration

FIGURE 6

Distribution of Left- and Right-leaning X accounts that authored posts for which:

- (A) Community Notes were requested;
- (B) Community Notes were proposed; and
- (C) Community Notes reached Helpful Status.

Since those statistics were computed over non-deleted posts with inferred ideological observed differences could be partially explained by differential deletion patterns between Left-and Right-leaning users.



Source: Authors' own elaboration

CONCLUSIONS

The ECAT Workshop 2025 underscored that children and young people face a wide range of platform-related risks. Addictive design features, cyberbullying and exposure to self-harm or eating disorder content intersect with emerging challenges such as misogynistic influencers and reliance on generative AI.

Across workshop sessions, speakers highlighted the need for stronger content moderation, age verification and design transparency practices, as well as coordinated mitigation strategies involving platforms, families, schools and regulators.

Experts also called for more data access and research to understand the impact of platforms, and through the abstracts presented in poster sessions and gathered in this publication, it is clear that researchers across disciplines are working to provide exactly this.

The protection of minors remains high on the research and policy support agenda of ECAT. The team will continue to play its part in making the online world a positive place for children and young people to engage and express themselves while being protected from harm.



LIST OF FIGURES

Figure 1: Overview of experiment methodology and architecture.....	14
Figure 2: Volume of content (including but not limited to reels, videos, static images and textual data) removed for 'sexual content and nudity' and 'violence and threats' across major social media platforms.....	17
Figure 3: Example item from the manosphere engagement task.....	31
Figure 4: Daily moderation volume, TikTok vs. Amazon.....	38
Figure 5: Community Notes usage and outcomes.....	44
Figure 6: Distribution of Left- and Right-leaning X accounts that authored posts for which (A) Community Notes were requested; (B) Community Notes were proposed; and (C) Community Notes reached Helpful Status.....	45

LIST OF TABLES

Table 1: Integrated Harm Framework with final voting statistics and key points of discussion.....22

Science for policy

The Joint Research Centre (JRC) provides independent, evidence-based knowledge and science, supporting EU policies to positively impact society



Scan the QR code to visit:

[The Joint Research Centre](https://joint-research-centre.ec.europa.eu)

<https://joint-research-centre.ec.europa.eu>

