Algorithmic Transparency: The EU Digital Services Act and Young People's Experiences of Online Platforms

MEGAN NYHAN*, University College Dublin, Ireland
PRANAV NARULA*, University College Dublin, Ireland
IZZY FOX, Dublin City University, Ireland
KEVIN DOHERTY, University College Dublin, Ireland
RUIHAI DONG, University College Dublin, Ireland
BARRY O'SULLIVAN, University College Cork, Ireland
JOSEPHINE GRIFFITH, University of Galway, Ireland
SUSAN LEAVY, University College Dublin, Ireland

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 Instagram, 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.

Keywords: Recommender Systems, Moderated Content, Digital Services Act, Trustworthy AI

Reference Format:
Megan Nyhan, Pranav Narula, Izzy Fox, Kevin Doherty, Ruihai Dong, Barry O'Sullivan, Josephine Griffith, and Susan Leavy. 2025. Algorithmic Transparency: The EU Digital Services Act and Young People's Experiences of Online Platforms. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 5 pages.

*Both authors contributed equally to this research.

Authors' Contact Information: Megan Nyhan, University College Dublin, Dublin, Ireland, megan.nyhan@ucdconnect.ie; Pranav Narula, University College Dublin, Dublin, Ireland, pranav.narula@ucdconnect.ie; Izzy Fox, Dublin City University, Dublin, Ireland, Izzy.Fox@mu.ie; Kevin Doherty, University College Dublin, Dublin, Ireland, kevin.doherty@ucd.ie; Ruihai Dong, University College Dublin, Dublin, Ireland, ruihai.dong@ucd.ie; Barry O'Sullivan, University College Cork, Cork, Ireland, b.osullivan@cs.ucc.ie; Josephine Griffith, University of Galway, Galway, Ireland, josephine.griffith@universityofgalway.ie; Susan Leavy, University College Dublin, Dublin, Ireland, susan.leavy@ucd.ie.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

1 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 [5, 4]. 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 [3, 1]. 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 their appropriateness and effectiveness of their moderation approcahes.

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 behavior (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 peoples' 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.

2 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 periods 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 "Non-Consensual Nudity", despite broadly covering overlapping issues. In our analysis, we found these inconsistencies

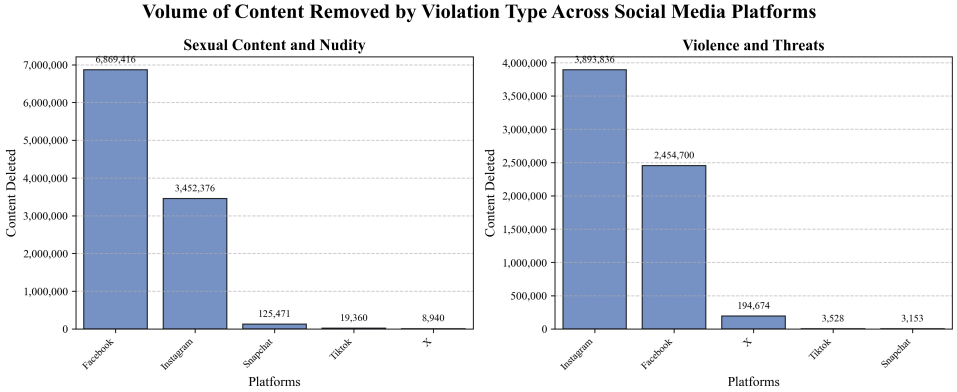

Fig. 1. Volume of Content (including but not limited to reels, videos, static images and textual data) Removed for "Sexual Content and Nudity" and "Violence and threats" Violations across Major Social Media Platforms.

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 peoples' 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. 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 (see Fig. 1), we focused our analysis on these categories. 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.

3 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 addressing 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 [2], encourages deeper reflection on the risks young users face online.

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

4 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.

5 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

[1] Matthew F Carlin. "Real Harm to Real People: A Restorative Justice Theory for Social Media Accountability". In: *N. Ky. L. Rev.* 51 (2024), p. 145.

[2] Mahmoud Kaddoura. "Think pair share: A teaching learning strategy to enhance students' critical thinking." In: *Educational research quarterly* 36.4 (2013), pp. 3–24.

[3] Carol F Scott et al. "Trauma-Informed Social Media: Towards Solutions for Reducing and Healing Online Harm". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581512. URL: https://doi.org/10.1145/3544548.3581512.

[4] Andreas Veglis. "Moderation Techniques for Social Media Content". In: *Social Computing and Social Media*. Ed. by Gabriele Meiselwitz. Cham: Springer International Publishing, 2014, pp. 137–148. ISBN: 978-3-319-07632-4.

[5] Sarah Myers West. "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms". In: *New Media & Society* 20.11 (2018), pp. 4366–4383. DOI: 10.1177/1461444818773059. eprint: https://doi.org/10.1177/1461444818773059. URL: https://doi.org/10.1177/1461444818773059.