EU DSA Election Guidelines Input

Prepared by Ruth Elisabeth Appel, PhD Candidate in Political Communication at Stanford University
March 6, 2024

Q1: Are there any documents, reports, guidelines, academic studies or relevant independent research you recommend as further input for these guidelines?

See "Related resources" sections in the responses to the questions below for further input suggestions.

Q2: How can the Commission further clarify the purpose and scope of these guidelines to better address systemic risks in electoral processes?

Q3: Do you agree with the recommended best practices in this section?

Article (16) lists specific mitigation measures to address systemic risks to electoral processes. While all of these measures are supported by the academic literature, it is important to note that e.g. suggestions (a) through (c) are focused on individual-level interventions aimed at equipping individuals to e.g. deal with misinformation, while other suggestions like (g) and (h) address the platform ecosystem.
Existing academic work focuses primarily on individual-level interventions ("i-frame") as opposed to system-level interventions ("s-frame"), likely because system-level interventions are extremely difficult to evaluate for external researchers without platform access.
S-frame interventions lift the burden off the individual, which may be conducive to effective interventions in complex settings like the social media environment (Acquisti et al., 2015, make this point related to privacy protection). While i-frame interventions are important, more effort and resources should be dedicated to investigating s-frame interventions (Chater & Loewenstein, 2023, Roozenbeek & Zollo, 2022), and to comparing different kinds of interventions at scale.

Article (16) d. addresses the moderation of virality of content that threatens the integrity of the electoral process. Existing research, e.g. as part of the Facebook and Instagram Election Study (FIES), shows that changing platform features – such as reducing exposure to reshared content – can indeed reduce the amount of content form untrustworthy sources that users are exposed to, but this can also have costs such as reducing exposure to political news (Guess et al., 2023). Therefore, further research and transparency on the effect of platform features is essential to design the most effective interventions.

Related resources:

- Chater, N., & Loewenstein, G. (2023). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X22002023
- Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, *347*(6221), 509–514. https://doi.org/10.1126/science.aaa1465
- Roozenbeek, J., & Zollo, F. (2022). Democratize social-media research — with access and funding. *Nature*, *612*, 404. https://doi.org/10.1038/d41586-022-04407-8

- Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., … Tucker, J. A. (2023). Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, *381*(6656), 404–408. https://doi.org/10.1126/science.add8424

Article (18) addresses "Third party scrutiny and research," an essential step towards gaining a better understanding of how elections are impacted by technology. The current draft provides only vague guidance. Existing academic-industry collaborations such as the Facebook and Instagram Election Study (FIES) that investigate the impact of Facebook and Instagram on the 2020 US elections could provide important insights on how to structure third party research, and what obstacles need to be taken into account.
For example, encouraging platforms to provide *sufficient resources* and *dedicated staff*, and to facilitate *timely analysis* (e.g., encouraged by publicized project timelines or dashboards) to inform public policy and science as soon as possible, would be important steps to ensuring that external researchers can inform election process integrity using platform data.
I agree with the importance of transparency and data access as emphasized by the Stanford Internet Observatory, e.g. in their comments on DSA Article 40.
In order for third party scrutiny to be effective, it is important that external researchers are given the required data access and protection. In the past, researchers have faced threats from individuals, companies (see e.g., Hatmaker, 2021) or political actors (see e.g., Myers & Frenkel, 2023) for engaging in misinformation research. Institutional protection for researchers is necessary to enable their work.

Related resources:

- Wagner, M. W. (2023). Independence by permission. *Science*, *381*(6656), 388–391. https://doi.org/10.1126/science.adi2430
- First FIES results: Science, 381, https://www.science.org/toc/science/381/6656; https://www.nature.com/articles/s41586-023-06297-w
- FIES project data repository: https://socialmediaarchive.org/search?cc=US2020&ln=en&c=US2020
- FIES project information: https://medium.com/@2020_election_research_project
- Hatmaker, T. (2021). Facebook cuts off NYU researcher access, prompting rebuke from lawmakers. https://techcrunch.com/2021/08/04/facebook-ad-observatory-nyu-researchers/
- Myers, S. L., & Frenkel, S. (2023, June 19). G.O.P. Targets Researchers Who Study Disinformation Ahead of 2024 Election. *The New York Times*. https://www.nytimes.com/2023/06/19/technology/gop-disinformation-researchers-2024-election.html

Q4: What additional factors should be taken into account by providers of VLOPs and VLOSEs when detecting systemic risks related to electoral processes??

Article (16) g. addresses demonetization of disinformation content, and h. encourages timely detection and removal of coordinated inauthentic manipulation.
Most existing research focuses on deceptive campaigns that are politically-motivated. Various reports about such campaigns are available, e.g. on Meta's Coordinated Inauthentic Behavior Archive or reports from organizations such as the Stanford Internet Observatory, Graphika or DFRLab.

However, there are financially-motivated organizations that are using politics as a lure, given their goal of creating highly enticing content (see e.g. Subramanian, 2017, Hughes, 2021). Current research and current platform policies and enforcement do not take such financially-motivated campaigns as seriously as politically-motivated coordinated inauthentic behavior – they often do not not even report on them –, even though the content produced and audiences reached might be similar. Actors need not be foreign or politically-motivated to cause harm. It is critical that financially-motivated campaigns are scrutinized by researchers, platforms and policymakers alike and not disregarded as harmless spam before their impact has been investigated.

In terms of integrity of services and removing manipulative content, ensuring regular reporting (on both politically- and financially-motivated campaigns) as well as transparency on how such campaigns are defined and detected could help create more clarity and accountability. The procedures platforms use for detection are obscure from the outside, and even for external academics collaborating with companies. While some details on detection might be sensitive and should not be disclosed to avoid helping threat actors, more general information about approaches and procedures to detect manipulative content would allow for constructive feedback on these processes and better coordination among companies, academics, civil society and government when it comes to combating election interference. To keep track of potential threats and assess evidence accumulated on both threat actors and the tactics they use, a centralized database on threats, such as the threat atlas "Adversarial Threat Landscape for Artificial-Intelligence Systems" that Microsoft is contributing to (Microsoft, 2024), as well as on threat actors, could be an important first step.

Related resources:

- Meta Coordinated Inauthentic Behavior Archive: https://about.fb.com/news/tag/coordinated-inauthentic-behavior/
- Hughes, H. C., & Waismel-Manor, I. (2021). The Macedonian Fake News Industry and the 2016 US Election. *PS - Political Science and Politics*, *54*(1), 19–23. https://doi.org/10.1017/S1049096520000992
- Subramanian, S. (2017, February). Inside the Macedonian Fake-News Complex. *WIRED*. https://www.wired.com/2017/02/veles-macedonia-fake-news/
- Microsoft. (2024). Staying ahead of threat actors in the age of AI. https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/

Q5: Are there additional mitigation measures to be considered as best practices on the basis of their proven effectiveness mitigating risks to electoral processes?

Q6: How should providers of VLOPs and VLOSEs measure effectiveness of their risk mitigation measures in a reliable and conceptually valid way for electoral processes?

Q7: Do you agree with the recommended best practices in this section?

Article (27) encourages red-teaming exercises, but does not address what is needed in terms of context and resources to make them successful.

Work by non-profit organizations (e.g., Friedler et al., 2023) and companies (e.g., Anthropic, 2023) provides more detailed insights on when and how red-teaming may work well. This includes a clear definition of the goal of the exercise, and seeing red-teaming as just one part of a broader process (Friedler et al., 2023). It also includes ensuring that information exchange with relevant e.g. government bodies is possible, that there are legal safeguards in place when red-teaming involves eliciting sensitive information, and that the process is standardized as much as possible for the sake of robustness and comparability (Anthropic, 2023).

Because companies try to protect the privacy of their red-team members, they often cannot provide, or do not know, detailed information about the red team members. Given that reinforcement learning based on human (red-teaming) feedback may introduce bias, e.g. more skewed political views, into generative AI systems (see e.g. Perez et al., 2022), it is important to make sure that red team members do not only have expertise, but are also diverse in their composition to ensure that no unintentional bias is introduced because further training of a model relies disproportionately on specific, non-representative groups.

Related resources:

- Friedler, S., Singh, R., Blili-Hamelin, B., Metcalf, J., & Chen, B. J. (2023). *AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability*. https://datasociety.net/wp-content/uploads/2023/10/Recommendations-for-Using-Red-Teaming-for-AI-Accountability-PolicyBrief.pdf
- Anthropic. (2023). Challenges in evaluating AI systems. https://www.anthropic.com/news/evaluating-ai-systems
- Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., … Kaplan, J. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. http://arxiv.org/abs/2212.09251

Q8: Which risks of generative AI for electoral processes should additionally be considered in this section?

Q9: What additional evidence-based best practices on risk mitigation for electoral processes related to the creation of generative AI content should be considered?

Q10: What additional evidence-based best practices on risk mitigation for electoral processes related to the dissemination of generative AI content should be considered?

Q11: What are best practices for providers of VLOPs and VLOSEs to ensure that their risk mitigation measures keep up with technological developments and progress?

Q15: Do you agree with the recommended best practices in this section?
Q16: Are there any additional measures that providers of VLOPs and VLOSEs should take specifically during an electoral period?

Q17: How can rapid response mechanisms be improved for handling election-related incidents on VLOPs or VLOSEs?

Q18: What other mechanisms should be considered to foster more effective collaboration with national authorities and civil society organizations?

Q19: Are there any additional resources that help providers of VLOPS and VLOSEs identify relevant organisations/experts at the national level?

Q20: Do you agree with the recommended best practices in this section?

Q21: What elements should be included in voluntary post-election review by providers of VLOPs or VLOSEs to assess the effectiveness of their risk mitigation strategies?

Q22: What are your views on the best practices proposed in this section?

Q23: What additional mitigation measures should be considered for the elections for the European Parliament present for online platforms?

Q24: What additional feedback or suggestions do you have regarding these guidelines?