Code of Practice on Disinformation – Report of AI Forensics for the period from April 2024 to the end of July 2024

Executive summary

Executive summary (max. 2 pages)

In the lead-up to the 2024 European elections, our organization conducted extensive research and analysis on the influence of emerging technologies on electoral integrity. We closely monitored the use of generative AI in political campaigns, identifying instances where political parties like Rassemblement National, Reconquête, and Les Patriotes deployed AI-generated imagery to manipulate public opinion and amplify misleading narratives. Our efforts extended to analyzing social media platforms like TikTok, where algorithmic features such as search suggestions were found to promote biased or misleading content. We also examined the role of chatbots in electoral contexts, highlighting the risks posed by misinformation and the need for more consistent, transparent, and accountable content moderation across platforms. Through these activities, we aimed to raise awareness of the threats to democratic processes and advocate for stronger regulatory measures to safeguard election integrity.

Guidelines for filling out the report

Crisis and elections reporting template

Relevant signatories are asked to provide proportionate and appropriate information and data during a period of crisis and during an election. Reporting is a part of a special chapter at the end of the harmonised reporting template and should follow the guidelines:

- The reporting of signatories' actions should be as specific to the particular crisis or election reported on as possible. To this extent, the rows on "Specific Action[s]" should be filled in with actions that are either put in place specifically for a particular event (for example a media literacy campaign on disinformation related to the Ukraine war, an information panel for the European elections), or to explain in more detail how an action that forms part of the service's general approach to implementing the Code is implemented in the specific context of the crisis or election reported on (for example, what types of narratives in a particular election/crisis would fall into scope of a particular policy of the service, what forms of advertising are ineligible).
- Signatories who are not offering very large online platform services and who follow the invitation to report on their specific actions for a particular election or crisis may adapt the reporting template as follows:
    - They may remove the "Policies and Terms and Conditions" section of the template, or use it to report on any important changes in their internal rules applicable to a particular election or crisis (for example, a change in editorial guidelines for fact-checkers specific to the particular election or crisis)
    - They may remove any Chapter Section of the Reporting Template (Scrutiny of Ads Placement, Political Advertising, Integrity of Services etc.) that is not relevant to their activities
- The harmonised reporting template should be filled in by adding additional rows for each item reported on. This means that rather than combined/bulk reporting such as "Depending on severity of violation, we demote or remove content based on policies X, Y, Z", there should be individual rows stating for example "Under Policy X, content is demoted or removed based on severity", "Under Policy Y, content [...]" etc.
- The rows should be colour-coded to indicate which service is being reported on, using the same colour code as for the overall harmonised reporting template.

Reporting should be brief and to the point, with a suggested character limit entry of 2000 characters.

Uploading data to the Transparency Centre

The reports should be submitted to the Commission in the form of the pdf via e-mail to the address CNECT COP TASK FORCE CNECT-COP-TASK-FORCE@ec.europa.eu within the agreed deadline. Signatories will upload all data from the harmonised reporting template to the Transparency Centre, allowing easy data access and filtering within the agreed deadline. It is the responsibility of the signatories to ensure that the uploading takes place and is executed on time. Signatories are also responsible to ensure that the Transparency Centre is operational and functional by the time of the reports' submission that the data from the reports are uploaded and made accessible in the Transparency Centre within the above deadline, and that users are able to read, search, filer and download data as needed in a user-friendly way and format.

Reporting on the signatory's response during an election

Reporting on the signatory's response during an election

2024 European Parliament Elections

Threats observed during the electoral period: [suggested character limit 2000 characters].
AI Forensics has been actively involved in election monitoring in 2024, with the following reports published as its outcomes:

1. **French Elections** (Artificial Elections: Exposing the Use of Generative AI Imagery in the Political Campaigns of the 2024 French Elections)
   AI Forensics investigated how AI-generated images were used in French political campaigns during the 2024 European Parliament and legislative elections. In May and June of 2024, we collected data from a variety of sources to get a comprehensive look at the use of AI imagery. We explored official party websites and their social media accounts on platforms such as Facebook, Instagram, X (formerly Twitter), TikTok, YouTube, and LinkedIn.

   Main threats:

   The lack of **transparency** is alarming and highlights several critical concerns. Firstly, political parties and social media platforms are failing to adequately disclose the use of AI-generated imagery, which undermines public trust. Additionally, there is a pressing need for **stricter content labelling** to ensure the integrity of political campaigns and prevent the spread of **misleading information**. Finally, our findings underscore the **necessity of reinforcing EU-wide policies on the use of generative AI** in elections to safeguard democratic processes and maintain electoral integrity.

2. **TikTok Search**: Analyzing TikTok´s "Others searched for" Feature: TikTok's impact on public discourse among young users in Germany, focusing on the influence of search suggestions. This investigation on TikTok "Others searched for"; feature helps to understand its influence on political discourse, especially in the context of the 2024 elections. Conducted in collaboration with AI Forensics and interface TikTok Audit Team, this study aimed to determine if TikTok´s algorithm promotes misleading or sensational content. This feature suggests search terms to users, which could potentially lead them to questionable information or politically biased content, posing significant risks to public discourse.

   Main threats:
   The study highlights that TikTok's "Others Searched For" feature **can distort reality for young users, especially during critical electoral periods**. This distortion can negatively affect public political discourse, making it imperative for social media platforms **to implement more robust oversight and transparency on their algorithms,** including on less prominent algorithmic features such as search suggestions.. Our findings emphasize the need for improved measures to ensure that search suggestions **do not perpetuate misinformation or political bias**, thus contributing to a more informed and balanced media environment.

3. **Chatbot (s)elected moderation**: Measuring the Moderation of Election-Related Content Across Chatbots, Languages and Electoral Contexts

   This report evaluates and compares the effectiveness of these safeguards in different scenarios. In particular, we investigate the consistency with which electoral moderation is triggered, depending on (i) the chatbot, (ii) the language of the prompt, (iii) the electoral context, and (iv) the interface.

   Main threats: The effectiveness of the **moderation safeguards deployed** by Copilot, ChatGPT, and Gemini is **widely different**. Gemini's moderation was the most consistent, with a moderation rate of 98%. For the same sample on Copilot, the rate was around 50%, while on the OpenAI web version of ChatGPT, there is no additional election-related moderation**. Moderation is strictest in English and highly inconsistent across languages**. When prompting Copilot about EU Elections, the moderation rate was the highest for English (90%), followed by Polish (80%), Italian (74%), and French (72%). It falls below 30% for Romanian, Swedish, Greek, or Dutch, and even for German (28%) despite it being the EU's second most spoken language. For a given language, when asking the analogous prompts for both the EU and the US

elections, **the moderation rate can vary substantially**. This confirms the inconsistency of the process. **Moderation is inconsistent between the web and API versions**. The electoral safeguards on the web version of Gemini have not been implemented on the API version of the same tool.

4. **No Embargo in Sight:** Meta leds pro-Russian propaganda flood the EU: This investigation sheds light on a **significant loophole in the moderation of political advertisements** on Meta platforms, highlighting systemic failures just as the European Union heads into crucial parliamentary elections. Our findings uncover a sprawling pro-Russian influence operation that **exploits these moderation failures, risking the integrity of democratic processes in Europe**.

   Main threats: Widespread Non-compliance: Less than 5% of undeclared political ads are caught by Meta's moderation system.**Ineffective Moderation**: 60% of ads moderated by Meta do not adhere to their own guidelines concerning political advertising. **Significant Reach**: A specific pro-Russian propaganda campaign reached over 38 million users in France and Germany, with most ads not being identified as political in a timely manner. **Rapid Adaptation**: The influence operation has adeptly adjusted its messaging to major geopolitical events to further its narratives.

Mitigations in place during the electoral period: [suggested character limit: 2000 characters].

[Note: Signatories are requested to provide information relevant to their particular response to the threats and challenges they observed on their service(s). They ensure that the information below provides an accurate and complete report of their relevant actions. As operational responses to crisis/election situations can vary from service to service, an absence of information should not be considered a priori a shortfall in the way a particular service has responded. Impact metrics are accurate to the best of signatories' abilities to measure them].

Policies and Terms and Conditions

Outline any changes to your policies

Policy	**Changes (such as newly introduced policies, edits, adaptation in scope or implementation)**	**Rationale**
		Our analysis on the French elections highlights several areas where policies and terms and conditions should respond to emerging threats related to generative AI in political campaigns: 1. **Transparency Requirements**: There is a critical need for greater transparency from political parties and social media platforms regarding the use of AI-generated imagery. Current policies must enforce clear disclosure when synthetic content is used in campaigns, ensuring the public is fully informed about

		AI-altered visuals. This should include a requirement for political actors to label AI-generated materials and for platforms to flag such content when shared on social media. 2. **Stricter Content Labelling**: To combat the spread of misleading or deceptive AI-generated content, platforms must enhance their content moderation policies. Automated tools and human oversight should work in tandem to identify and remove manipulated or misleading images that distort political discourse. Policies should also include stringent checks to ensure that AI-generated content used in political contexts complies with electoral laws and ethical standards. 3. **Translating Codes of Conduct into regulatory obligations**: The findings underline the necessity of strengthening EU-wide policies on the use of generative AI in elections. Current frameworks, like the Code of Conduct for the 2024 European Parliamentary Elections, should be reinforced with mandatory regulations, penalties for violations, and robust enforcement mechanisms. This will safeguard democratic processes from the undue influence of misleading, AI-generated content and maintain electoral integrity across member states. 4. **Amplification of Misinformation**: Generative AI has been used to produce content that spreads misinformation, emotionally manipulates voters, and supports extremist ideologies. The ease and low cost of creating such content exacerbate the risk of misleading narratives dominating electoral campaigns. Our report on TikTok´s "Others Searched for" Feature suggests several solutions to address the threats: 1. **Stronger Oversight to prevent algorithmic harms**: Social media platforms, especially TikTok, should strengthen their content moderation systems to prevent misleading or biased search suggestions. This includes actively identifying and removing dog whistles, misinformation, and content designed to manipulate users' political views. 2. **Transparency in Algorithms**: Platforms must be more transparent about how their algorithms generate search suggestions. Clear policies are needed to explain how suggestions are ranked, especially during election periods, to ensure that users aren't steered toward specific political narratives or parties. 3. **Reducing Political Bias**: TikTok should implement safeguards to ensure that search suggestions do not disproportionately promote one political party or viewpoint. By doing so, they can help foster a more balanced media environment that avoids distorting electoral discourse. Our report on "Chatbot (s)elected moderationbsuggests the following solutions to address the threats posed by chatbot moderation and misinformation in sensitive contexts such as elections: 1. **Consistency in Moderation**: Platforms must ensure that chatbot moderation mechanisms are applied uniformly across all languages and geographies, preventing gaps in protection for non-English users and elections in various regions. 2. **Transparency of Moderation Systems**: Platforms should publish clear documentation explaining the design, implementation, and functioning of their moderation systems, helping users and researchers understand how content is managed and ensuring safeguards are in place. 3. **Accountability through External Scrutiny**: Introducing research APIs that allow third parties to test and scrutinize chatbot moderation layers is essential for improving accountability. This would enable external experts to assess the effectiveness of the moderation mechanisms and identify potential biases or inconsistencies.

		4. **Improved Moderation for Sensitive Prompts**: Platforms should develop robust safeguards for sensitive topics, such as elections, ensuring that chatbots do not spread harmful misinformation or propaganda. Enhanced moderation must be implemented systematically across all contexts.
Scrutiny of Ads Placements		
Outline approaches pertinent to this chapter, highlighting similarities/commonalities and differences with regular enforcement.		
Specific Action applied (with reference to the Code's relevant Commitment and Measure)	Description of intervention	
	Indication of impact including relevant metrics when available	
Specific Action applied (with reference to the Code's relevant Commitment and Measure)	Description of intervention	
	Indication of impact including relevant metrics when available	
Political Advertising		
Outline approaches pertinent to this chapter, highlighting similarities/commonalities and differences with regular enforcement.		
Specific Action applied (with reference to the Code's	Description of intervention	

relevant Commitment and Measure)	The report “No Embargo in Sight” suggests several key solutions to address the threats to electoral integrity posed by online platforms ahead of the EU elections: 1. **Launch infringement proceedings** against Meta under the Digital Services Act (DSA) for systemic risks, emphasizing Meta's failure to address Coordinated Inauthentic Behavior that threatens election integrity. 2. **Enforce stricter application of DSA Article 39** to require platforms to provide comprehensive metadata in their ad registries, enabling external scrutiny of political ads. Platforms like X should improve transparency in line with Meta's standards. 3. **Immediate action by Meta** to neutralize the ongoing "Doppelgänger" influence operation and preemptively moderate any new similar activity. 4. **Automate the labelling of political ads** with systems to flag political content and enforce the necessary disclosure and targeting requirements, ensuring compliance with EU regulations.
	Indication of impact including relevant metrics when available
Integrity of Services	
Outline approaches pertinent to this chapter, highlighting similarities/commonalities and differences with regular enforcement.	
Specific Action applied (with reference to the Code’s relevant Commitment and Measure)	Description of intervention
	Indication of impact including relevant metrics when available
Empowering Users	
Outline approaches pertinent to this chapter, highlighting similarities/commonalities and differences with regular enforcement.	
Specific Action applied (with reference to the Code’s	Description of intervention Our report on TikTok´s “Others Searched for” Feature suggests several solutions to address the threats:

relevant Commitment and Measure)	**User Education**: Platforms should provide educational tools to help users critically assess the information they encounter, promoting media literacy and a deeper understanding of potential biases within search suggestions.
	Indication of impact including relevant metrics when available
Empowering the Research Community	
Outline approaches pertinent to this chapter, highlighting similarities/commonalities and differences with regular enforcement.	
Specific Action applied (with reference to the Code's relevant Commitment and Measure)	Description of intervention
	Indication of impact including relevant metrics when available
Empowering the Fact-Checking Community	
Outline approaches pertinent to this chapter, highlighting similarities/commonalities and differences with regular enforcement.	
Specific Action applied (with reference to the Code's relevant Commitment and Measure)	Description of intervention Our report on TikTok´s "Others Searched for" Feature suggests several solutions to address the threats: **Fact-Checking and Flagging of Sensitive Content**: Implementing robust fact-checking mechanisms that flag potentially misleading or biased search suggestions would help young users navigate political content more responsibly.
	Indication of impact including relevant metrics when available