College of Europe
Collège d'Europe
Brugge
Natolin

NATOLIN POLICY PAPERS SERIES

Mitigating Disinformation Risks in the Run-Up to 2024 European Elections: A Three-Step Action Plan against AI-Generated Disinformation

by Anna Denais

mentored by: **Olaf Osica**

NATOLIN
DIGITAL TRANSFORMATION
NEST

VOL. 1/2024

Natolin Policy Papers Series vol. 1/2024

Mitigating Disinformation Risks in the Run-Up to 2024 European Elections: A Three-Step Action Plan against AI-Generated Disinformation

Author:	Anna Denais
Volume Editor:	Barbara Bobrowicz
Mentored By:	Olaf Osica
Publication Coordinator:	Álvaro Garrote Fuentes
Design And Typesetting:	Maja Grodzicka
Proofread by:	Mateusz Byrski

Published In Poland By:

Brugge College of Europe Collège d'Europe Natolin

College of Europe in Natolin
www.coleurope.eu
www.natolin.eu

84, Nowoursynowska Str., 02-797 Warsaw, Poland

ISBN 978-83-63128-12-8	ebook
ISSN	Natolin Policy Papers Series

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

EXECUTIVE SUMMARY

The escalating threat of AI-generated disinformation poses significant challenges to the integrity of elections and democratic processes. The existing legislative mechanisms lack specific provisions to combat AI-generated disinformation effectively. Yet, given the hyperrealistic nature of such disinformation and thus its potential to influence public opinion, there is an urgent need for targeted interventions.

This policy paper proposes an **Action Plan directed towards the European Commission** to address AI-generated disinformation comprehensively. Building upon existing legislation, the Action Plan outlines three key steps:

1. **Minimizing the production of AI-generated disinformation** by incentivizing companies behind AI tools to implement safeguards in these;

2. **Mitigating its dissemination on online platforms by labelling AI-generated content** through watermarking techniques and to enhance users' awareness;

3. **Establish a framework for exchanges between AI tools and online platforms** to facilitate cooperation, share best practices and develop common standards.

Furthermore, this policy paper suggests the potential transformation of this Action Plan into a Code of Conduct under the frameworks of the Digital Services Act (DSA) and the Artificial Intelligence Act (AI Act), to provide stakeholders with incentives to combat AI-generated disinformation.

College of Europe
Collège d'Europe
Brugge
Natolin

NATOLIN POLICY PAPERS SERIES

Mitigating Disinformation Risks in the Run-Up to 2024 European Elections: A Three-Step Action Plan against AI-Generated Disinformation

by Anna Denais

TABLE OF CONTENTS

P.6–7

Introduction

The issue of AI-generated disinformation

The current framework

P.8–9

Impacts of AI-generated disinformation on elections

Case 1: Elections in Slovakia in 2023

Case 2: Upcoming elections in the United States in November 2024

Case 3: Upcoming European elections in June 2024

P.9–12

Policy recommendations

Action Plan against AI-generated disinformation

Step 1: Incentivize AI tools creators to prevent disinformation at the tool production stage

Step 2: Prioritize AI-labelling through watermarking as a mitigation strategy

Step 3: Create a framework for exchanges between AI tools and online platforms

P.13

The perspective of turning this Action Plan into a Code of Conduct: a way to incentivize stakeholders

P.14

Concluding remarks

P.14–15

Bibliography

Mitigating Disinformation Risks in the Run-Up to 2024 European Elections: A Three-Step Action Plan against AI-Generated Disinformation

INTRODUCTION

The issue of AI-generated disinformation

According to the World Economic Forum, AI-generated disinformation is "the world's biggest short-term threat". During its annual meet-ing in Davos (Jan 2024), the organization highlighted its concerns about the impact of the phenomenon on the disruption of politics and polarization of societies[1] through its declaration which spotlit the issue of AI-generated disinformation. Over the recent years, AI tools have been getting more and more sophisticated, widespread and accessible.[2] They enable the quick production of false content at a very low cost.[3] This poses risks in terms of disinformation, as the hyperrealistic nature of AI-generated content makes it difficult to differentiate reality from false content, such as deep fakes.[4]

This risk takes on even greater significance, as 2024 is regarded as the "ultimate election year", with 64 elections taking place around the world, including the elections for the European Parliament.[5]

The current framework

Currently, disinformation and AI are addressed separately through several legislations. First, the DSA deals with disinformation as a "systemic risk" and sets instruments to tackle it, such as obligations for platforms to conduct risk assessments and mitigation measures.[6] Second, the AI Act passed by the European Parliament in March 2024

[1] Larry Elliott, 'AI-driven misinformation "biggest short-term threat to global economy"' *The Guardian* (10 January 2024).

[2] Seb Butcher, '2024 may be the year online disinformation gets the better of us' *Politico* (10 March 2024).

[3] 'Looking Ahead Generative AI' Carnegie Endowment for International Peace (2024) <https://carnegieendowment.org/2024/01/31/looking-ahead-generative-ai-pub-91489> accessed 8 May 2024.

[4] Bernard Marr, 'Will All Content Soon Be Fake?' *Forbes* (20 March 2024).

[5] Ewe Koh, 'The Ultimate Election Year: All the Elections Around the World in 2024' *Time* (28 December 2023).

[6] 'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC' *Official Journal of the European Union* L 277 (27 October 2022) Articles 34 and 35.

sets up a framework to frame the developments and uses of AI. Yet, it does not address it in light of disinformation. There is only one pertinent provision which states that artificial content needs to be labelled as such.[7] This current, albeit too broad framework cannot effectively prevent the sprawl of AI-generated disinformation, despite the increasing importance of doing so. Only the Strengthened Code of Practice on Disinformation approaches the issue directly, by inviting online platforms to counter and detect manipulative practices from AI systems. The Code is however a voluntary tool that only applies to a set of voluntary signatories.[8]

The need for a specific instrument on AI-generated disinformation

As outlined above, there are currently no instruments directly addressing AI-generated disinformation in an obligatory manner and taking into account the specificities of this type of disinformation. It seems appropriate to devise an Action Plan to bridge this gap. In 2018, an Action Plan on disinformation was published in the approach to the 2019 elections of the European Parliament to mitigate the threat.[9]

This raises the question of updating such an instrument to include the most recent developments. Disinformation can undermine election integrity and have a substantial destructive impact on democracy. Thus, as the 2024 European elections are approaching, the role played by AI in amplifying and disguising disinformation must be tackled.

[7] 'Artificial Intelligence Act: MEPs adopt landmark law.' News European Parliament (13 March 2024) <https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law> accessed 8 May 2024.

[8] European Commission, 'EU strengthened Code of practice on Disinformation' (2022) <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation> accessed 8 May 2024.

[9] European Commission, 'Communication from the Commission. Action Plan against Disinformation' (Brussels, 5 December 2018), <https://commission.europa.eu/document/download/b654235c-f5f1-452d-8a8c-367e603af841_en?filename=eu-communication-disinformation-euco-05122018_en.pdf> accessed 25 March 2024.

IMPACTS OF AI-GENERATED DISINFORMATION ON ELECTIONS

Since the launch of ChatGPT in 2022, this threat has been increasing with the democratization of AI tools[10], as the following cases show.

Case 1: Elections in Slovakia in 2023

The parliamentary elections held in Slovakia in 2023 serve as an apt example where disinformation was spread online through AI-generated audios. Two days before the elections, a fake telephone conversation between the leader of a Slovakian political party, Michal Šimečka, and a journalist was shared. In the conversation, the politician spoke of ways to manipulate the elections in favour of his party, including through potential purchase of votes. Another fake audio targeted this politician, in which he explained his plan to increase the price of beer after the elections. Both pieces of AI-generated disinformation were later refuted by fact-checkers, yet in the meantime they circulated virally.[11]

Case 2: Upcoming elections in the United States in November 2024

The campaign period before the upcoming elections in the United States has also been fallen prey to AI-generated disinformation. For instance, a false video was released online in late 2023 by the Republicans. It featured a dystopian world, which would result from Joe Biden's second term in office, with migrants flooding the country and a collapse of the financial system. The incumbent American president was also targeted by an AI-generated audio, a false telephone conversation in which he encouraged citizens not to vote in the elections.[12]

Case 3: Upcoming European elections in June 2024

The phenomenon has affected some candidates in the run-up to the elections of the European Parliament in June 2024. Many cases of AI-generated disinformation have already surfaced online. Among them, there is an AI-generated image of the former Commission-

[10] Nivedita Hazra, 'AI elections: Voting in the era of deepfakes and disinformation' *Firstpost* (6 May 2024).

[11] 'Prebunking AI-generated disinformation ahead of EU elections' European Digital Media Observatory (2024) <https://edmo.eu/publications/prebunking-ai-generated-disinformation-ahead-of-eu-elections/> accessed May 8, 2024.

[12] Hazra (n 10).

er for Climate Action Frans Timmermans in a private jet. There is also a false image of tractors and straw bales in front of the Eiffel Tower aiming to distort the public opinion's perception of the scale of farmers' protests.[13] It thus seems evident that the key issues for the elections have been instrumentalized for disinformation purposes.

According to the European Digital Media Observatory (EDMO), this phenomenon could expand further as the elections draw near. In particular, it could be used to discredit politicians, spread false narratives and undermine the integrity of the process. AI-generated disinformation thus has the potential to affect the European elections. As mentioned previously, this type of disinformation can take many shapes, with AI-generated videos, audio, pictures, and texts. Currently, the EDMO sees deep-fake audio as the most prominent risk due to its quality and, to a lesser extent, video and image. Nevertheless, it still warns of the necessity to follow the developments of all AI-generated content, including text, as technological progress could result in better quality and thus better ability to deceive European voters.[14]

POLICY RECOMMENDATIONS

The cases mentioned highlight the various forms AI-generated disinformation takes and its ability to sprawl by taking advantage of the ubiquity and universality of social media where it lives and spawns.

Action Plan against AI-generated disinformation

The following three-step Action Plan on AI-generated disinformation is recommended to tackle the issue by focusing on and leveraging the specificities of the phenomenon.

1. **Minimize the production of AI-generated disinformation** by blockage mechanisms to be employed at the production stage when necessary;

2. **Enforce labelling of AI-generated content** disseminated on online platforms through watermarking;

3. **And encourage exchanges between AI tools and online platforms** to increase their capacities to prevent disinformation with regard to the production and dissemination stages.

[13] 'Prebunking AI-generated disinformation ahead of EU elections.' (n 11)

[14] Ibid.

This Action Plan's effectiveness relies on its ability to combat the issue at the two initial stages – creation and dissemination. Indeed, as seen previously, AI significantly increases the amount of disinformation present online. This makes traditional methods such as fact-checking more difficult to implement, as fact-checkers are faced with an increasing workload. Moreover, AI makes disinformation more and more difficult to detect for users. For instance, fake AI-generated audio is very difficult to detect by social media's end users.[15] This is why it is so widely used in the context of elections to deceive voters.

Step 1: Incentivize AI tools creators to prevent disinformation at the tool production stage

The first step of the Action Plan to limit AI-generated disinformation would be to **act at its roots, before its production**. Indeed, AI tools providers can decide to block their services, when detecting their abuse for disinformation purposes. Some AI tools creators/providers have started to explore this option,[16] however, as of now, circumventing the safeguards is relatively easy. This is what two researchers from an Australian university attempted to show by creating disinformation content about vaccines and vaping. In a short time, and with no previous knowledge of AI, they were able to find online help on how to bypass the limits set up by AI tools and create a large amount of disinformation content.[17] Therefore, it appears clear that it is necessary to reinforce the mechanisms setting boundaries on content production.

Such a measure makes sense in a context where the increase in AI-generated disinformation is linked to the democratization of AI tools. Any actor, with the intention to deceive voters, can do so using AI tools available online. This is evident from the examples cited above. Therefore, limiting such access would help **reduce the amount of AI-generated disinformation created.**

However, this measure may reduce but cannot guarantee to block the use of AI services when an actor attempts to produce content for disinformation purposes.[18] Moreover, investing in this type of instrument could help improve its efficiency over time. **Effort should be invested in research and development** to increase the ability of AI tools to detect the intent of users to create disinformation content.

[15] Ibid.

[16] Nick Robins-Early, 'Disinformation reimagined: how AI could erode democracy in the 2024 US elections' *The Guardian* (19 July 2023).

[17] Melissa Davey, '"Alarming": convincing AI vaccine and vaping disinformation generated by Australian Researchers' *The Guardian* (13 November 2023).

[18] Robins-Early (n 16).

Step 2: Prioritize AI-labelling through watermarking as a mitigation strategy

The second step of the Action Plan would involve compulsory labelling of AI-generated content. This takes place **during the phase of dissemination through online platforms**. As a matter of fact, online platforms play a crucial role in exacerbating disinformation. They are the key channels through which disinformation reaches a large public. They also amplify this type of content, as it generates substantial engagement and platforms use algorithms promoting engagement.[19] Acting at this stage could then be beneficial in reducing AI-generated disinformation that may not be detected at its production stage involving AI tools.

In particular, labelling AI-generated content provides for **boosting social media users information literacy allowing them to detect** potential use of content for disinformation purposes. As seen in previous examples, disinformation content is mostly shared on social media where it can reach a large population of voters and fact-checking this type of content requires time which makes preventing its proliferation more difficult. This is well illustrated by Case 1 described above where some AI-generated fake audio was massively shared before later being refuted by experts during the 2023 Slovakian elections. This highlights the need for quicker and automated actions to avoid such dissemination of disinformation.

In that regard, labelling AI-generated content would appear to be particularly useful. This measure means that when content has been generated with AI, there is a visible warning that signals it. This makes users aware that the content is not real and might be used for disinformation purposes. and thus voters can approach such content with caution effectively preventing falling prey to deception.

To ensure that this labelling is efficient, **watermarking should be promoted as a mitigation strategy**. Watermarking takes place during the production of content in AI systems. A watermark invisible to users is then put on the content. This watermark is however visible to algorithms, which can trace it back to the AI model.[20] Therefore, once uploaded online, the watermark can allow online platforms to automatically set a label and warn users.

[19] Mario Mariniello, *Digital Economic Policy: The Economics of Digital Markets from a European Union Perspective* (Oxford University Press, 2022) 341-349.

[20] 'Generative AI and watermarking' European Parliament Research Service (2023) <https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf> accessed 8 May 2024.

Step 3: Create a framework for exchanges between AI tools and online platforms

Finally, to increase the actorness of AI tools creators/providers and online platforms, the Action Plan should lead the creation of a framework for exchanges. Through discussions and sharing of best practices, **all stakeholders could reinforce their capabilities** to counter AI-generated disinformation.

A first benefit of this multi-way discussion framework would be capacity building to **develop common standards for the compatibility of watermarking techniques**. Currently, there are no standardized watermarking techniques, which makes it difficult for online platforms to detect it in their algorithms.[21] Cooperation between AI tools creators/providers and platforms would enable development of one standard method or allow all algorithms to be able to detect all methods used thus leading to more efficient labelling of AI-generated content.

Secondly, such a framework would help AI tools creators/providers and online platforms to **follow the developments in AI-generated disinformation**. As the field evolves quickly, sharing identified new developments could be beneficial to other stakeholders to re-adapt their countermeasures. For instance, in the event that disinformation campaigns are increasing on a given topic, online platforms could communicate the need for AI tools creators/providers to pay increased attention to how their services are used to create content on this particular topic.

Moreover, such exchanges could encourage coordination to tackle emerging challenges. For instance, if one detects a way in which users avoid safeguards imposed on AI tools, a common solution may be found more quickly.

[21] Ibid.

THE PERSPECTIVE OF TURNING THIS ACTION PLAN INTO A CODE OF CONDUCT: A WAY TO INCENTIVIZE STAKEHOLDERS

In this current form, the Action Plan would not have a binding effect on the work of stakeholders. Rather, its value lies in the **creation of a roadmap customised to meet the specificities of AI-generated disinformation for its effective combating.**

Nevertheless, its impact could be enhanced by incentivizing stakeholders to take part in the framework. A means for that purpose would be to **transform it into a Code of Conduct** and thus changing the status of such regulations which is present in both the DSA and the AI Act. Online platforms are subject to the DSA, while AI tools are subject to the AI Act. Codes of Conduct are a means to guide the actions of stakeholders to address systemic risks and implement the legislation. For these companies, it is also a way to prove their compliance.[22] [23]

In this context, making this Action Plan a Code of Conduct would **encourage stakeholders to bolster their actions** to combat AI-generated disinformation. Indeed, it would provide them with incentives to show that they are complying with the DSA and the AI Act and avoid costly enforcement actions.

Moreover, it should be noted that this perspective aligns with the current work of the European Commission. As part of the enforcement of the DSA, it has already requested information from platforms on measures taken to limit the risks related to generative AI.[24] Therefore, this Action Plan, in the three steps described afore, would provide a response to these challenges and build on actions already undertaken by the Commission.

[22] European Parliament, 'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC' Article 35.

[23] 'Artificial Intelligence Act', Article 69 <https://artificialintelligenceact.com/title-ix/article-69/> accessed 8 May 2024.

[24] European Commission, 'Commission sends requests for information on generative AI risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines under the Digital Services Act' (14 March 2024) <https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very> accessed 25 March 2024.

CONCLUDING REMARKS

In conclusion, combating AI-generated disinformation demands tailored mechanisms and strategies. In light of current developments, it appears essential to develop more sophisticated instruments that can address its specificities. This Action Plan thus offers a possible avenue to tackle the phenomenon more effectively.

In that regard, the Action Plan emphasizes the need for concrete measures in two key phases: the creation process employing AI tools and the dissemination stage on online platforms. It also highlights the potential of fostering cooperation between stakeholders and developing common standards. As the phenomenon of AI-generated disinformation continues to grow, these solutions provide practical answers that will help mitigate its effects on future elections and thus defend democracy.

BIBLIOGRAPHY

1. 'Artificial Intelligence Act: MEPs adopt landmark law.' *News European Parliament* (13 March 2024) <https://www.europarl.europa.eu/news/en/pressroom/ 20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law> accessed 8 May 2024.

2. 'Artificial Intelligence Act', Article 69 <https://artificialintelligenceact.com/title-ix/article-69/> accessed 8 May 2024.

3. 'Generative AI and watermarking' European Parliament Research Service (2023) <https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/757583/EPRS_BRI(2023)757583_EN.pdf> accessed 8 May 2024.

4. 'Looking Ahead Generative AI' *Carnegie Endowment for International Peace* (2024) <https://carnegieendowment.org/2024/01/31/looking-ahead-generative-ai-pub-91489> accessed 8 May 2024.

5. 'Prebunking AI-generated disinformation ahead of EU elections.' *European Digital Media Observatory* (2024) <https://edmo.eu/publications/prebunking-ai-generated-disinformation-ahead-of-eu-elections/> accessed May 8, 2024.

6. 'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC.' *Official Journal of the European Union* L 277 (27 October 2022).

7. Butcher S, '2024 may be the year online disinformation gets the better of us' *Politico* (10 March 2024).

8. Davey M, '"Alarming": convincing AI vaccine and vaping disinformation generated by Australian Researchers' *The Guardian* (13 November 2023).

9. Elliott L, 'AI-driven misinformation "biggest short-term threat to global economy".' *The Guardian* (2024).

10. European Commission, 'Commission sends requests for information on generative AI risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines under the Digital Services Act' (14 March 2024) <https://digital-strategy.ec.europa.eu/en/news/commission-sends-requests-information-generative-ai-risks-6-very-large-online-platforms-and-2-very> accessed 25 March 2024.

11. European Commission, 'Communication from the Commission. Action Plan against Disinformation' (Brussels, 5 December 2018) <https://commission.europa.eu/document/download/b654235c-f5f1-452d-8a8c-367e603af841_en?filename=eu-communication-disinformation-euco-05122018_en.pdf> accessed 25 March 2024.

12. European Commission, 'EU strengthened Code of practice on Disinformation' (2022) <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation> accessed 8 May 2024.

13. European Parliament, 'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC' Article 35.

14. Hazra N, 'AI elections: Voting in the era of deepfakes and disinformation' *Firstpost* (6 May 2024).

15. Koh E, 'The Ultimate Election Year: All the Elections Around the World in 2024' *Time* (28 December 2023).

16. Mariniello M, *Digital Economic Policy: The Economics of Digital Markets from a European Union Perspective* (Oxford University Press, 2022).

17. Marr B, 'Will All Content Soon Be Fake?' *Forbes* (2024).

18. Robins-Early N, 'Disinformation reimagined: how AI could erode democracy in the 2024 US elections.' *The Guardian* (2023).

Brugge

College of Europe
Collège d'Europe

Natolin

NATOLIN POLICY PAPERS SERIES

College of Europe in Natolin
Nowoursynowska 84, 02-797 Warszawa, Polska

coleuropenatolin.eu