September 2023

PLATFORMS' POLICIES ON AI-MANIPULATED AND GENERATED MISINFORMATION

v1

EU DISINFO LAB

TABLE OF CONTENTS

Author: **Raquel Miguel**, EU DisinfoLab

Reviewer: **Noémie Krack**, KU Leuven Centre for IT & IP Law – imec

Layout and design: **Heini Järvinen**, EU DisinfoLab

EXECUTIVE SUMMARY

The development of artificial intelligence (AI) technologies has long been a challenge for the disinformation field, allowing content to be easily manipulated and contributing to accelerate its distribution. Focusing on content, recent technical developments, and the growing use of generative AI systems by end-users have exponentially increased these challenges, making it easier not just to modify but also to create fake texts, images, and audio pieces that can look real. Despite offering opportunities for legitimate purposes (e.g., art or satire), AI content is also widely generated and disseminated across the internet, causing – intentionally or not – harm and deception.

In view of these rapid changes, it is crucial to understand how platforms face the challenge of moderating AI-manipulated and AI-generated content that may end up circulating as mis- or disinformation. Are they able to distinguish legitimate uses from malign uses of such content? Do they see the risks embedded in AI as an accessory to disinformation strategies or copyright infringements, or consider it a matter on its own that deserves specific policies? Do they even mention AI in their moderation policies, and have they updated these policies since the emergence of generative AI to address this evolution?

Answers to these questions are crucial as the Digital Services Act (DSA) will provide new complaint mechanisms for users on the lack of enforcement of terms and conditions. The DSA will also require platforms to assess their mitigation measures (and results) against systemic risks.

The present factsheet delves into how some of the main platforms – Facebook, Instagram, TikTok, X (formerly Twitter), and YouTube – approach AI-manipulated or AI-generated content in their terms of use, exploring how they address its potential risk of becoming mis- and disinformation.

The analysis concluded that definitions are divergent, leaving users and regulators with diverse mitigation and resolution measures. First, only Facebook and TikTok mention "artificial intelligence" (including deepfakes in the case of Facebook) directly in their policies aiming to tackle disinformation. TikTok and X include "synthetic media" in their policies about manipulated and misleading media.[1] Therefore, it is not always possible to distinguish between the general misinformation policy and AI-specific considerations. Besides, the platforms often overlook mentioning AI-generated text and refer mainly to images and videos in their policies.

In cases, like TikTok, where platforms explicitly address synthetic or manipulated media with AI, they try to distinguish between allowed and banned uses. Little variations in the rationale behind content moderation exist: the driving force is either the misleading and harmful potential or a more compliance-oriented approach in terms of copyright and quality standards of the content.

On a different note, all the studied platforms qualify as Very Large Online Platforms (VLOPs) according to the DSA. The DSA is technically neutral, i.e., it applies regardless of the technology used to produce the content. Meanwhile, the strengthened Code of Practice on Disinformation has been reinforced by the co-regulatory mechanism and additional obligations to combat disinformation[2] set up by the DSA. In its 15th commitment, relevant signatories of the Code[3] are specifically called to "establish or confirm their policies in place for countering prohibited manipulative practices for AI systems that generate or manipulate content, such as warning users and proactively detect such content".

Consequently, even though X has withdrawn from the Code, it still has to abide by the DSA. Therefore, all the five studied platforms must comply with the DSA due diligence obligations and justify the means they deploy to combat disinformation on their services. This could require that they adopt new measures: among other required actions, platforms should update their policies to meet new needs in the face of rapidly evolving technologies, enhance cooperation with experts, and clarify the burden of responsibility on this complex topic.

1 Google mentions "synthetic content" in its recently announced policy related to political advertisements that presumably will apply to YouTube. However, due to its restricted scope, it cannot be considered a general policy for the entire video platform).

2 including through systemic risks assessment and mitigation, crisis protocols, users empowerment measures and increased transparency requirements.

3 All of the studied platforms except X.

PLATFORMS' POLICIES ON AI-MANIPULATED AND GENERATED MISINFORMATIVE CONTENT

EU DisinfoLab has developed an analytical framework to analyse and compare the policies of five platforms on different misinformative topics. Factsheets on electoral, health, and climate change misinformation have already been published following this framework. The same methodology (focusing on definitions and actions, and types of actions) is applied to AI-generated and manipulated misinformation. As far as applicable, the notes included in the table are verbatim mentions of the platforms' policies. In other cases, for the sake of simplification, the notes are a summary or analysis by the author.

CROSS-PLATFORM COMPARISON

Common Traits	Facebook	Instagram	YouTube	TikTok	X
Definition of synthetic/manipulated content	x		x	x	x
Mention of AI	x			x	x
Distinction between allowed and banned uses of manipulated or generated content (i.e., with AI)	x			x	x
Rationale for removing manipulated or generated content (i.e., with AI) based on risk of harm or to mislead	x		x	x	x
Specific AI resources	*				
Human content moderators	x	x	x	x	x
Automated moderation	x	x	x	x	x
Collaboration with experts	x		**	x	***
Collaboration with fact-checkers	x	x	***	***	**
Community contributions to content moderation	x	x	x		x
Labelling manipulated or generated content (i.e., with AI)	x	x	x	x	x
User responsibility in labelling or removal manipulated or generated content (i.e., with AI)				x	x
Downranking of manipulated or generated content (i.e., with AI)	x	x	x	x	x
Demonetisation of manipulated or generated content (i.e., with AI)	x	x	x	x	x
Strike policy	x	x	x	x	x
Removal of manipulated or generated content (i.e., with AI)	x	x	x	**	x
Prohibition of manipulated or generated content (i.e., with AI)	x	x	x	x	x
Advertising/monetisation standards for manipulated or generated content (i.e., with AI)	x	x	x	x	x
Misinformation policies updated in 2023	x		****	x	x

* Project Deepfake detection challenge
** Lack of clarity
*** Limited scope to specific countries
**** Limited scope to specific content (political ads) and specific countries

Disclaimer: In some cases, the platforms take a general approach and there are no specifications for AI-generated or manipulated content, but an 'x' is marked if the generic policies apply.

DEFINITIONS AND ACTORS

Platform	Definition of AI-manipulated or generated content and mention to AI	Distinction between allowed and banned AI-manipulated or generated content	Rationale for removing AI-manipulated or generated content	AI-related resources	Internal actors	External collaborators
Facebook	Manipulated media. Mentions AI, deepfakes, machine learning.	Banned: Content edited or synthesised – beyond adjustments for clarity or quality – in ways that would likely mislead. Product of AI or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic. Manipulated media that violates the platform's Community Standards. Allowed: Parody or satire, or video edited solely to omit or change the order of words.	High potential to mislead (and go viral quickly) Other misinformation policies refer to the risk of physical harm, or interference with political processes.	Project Deepfake detection challenge	Human and automated moderation, including AI technologies.	Third-party fact-checkers. Feedback from the community. Partnering with academia, government and industry.
Instagram	No mention of AI content, general approach regarding false information.	No	Violations of community guidelines.	None	Human and automated content moderation, including AI technologies.	Third-party fact-checkers. Feedback from the community.
YouTube	Manipulated content is mentioned as misleading or deceptive content, but no specifications about the technology used or mentions of AI. Synthetic content is mentioned in Google's Updates to Political content policy that presumably will apply to YouTube, but with a restricted scope (for political ads in limited regions).	Banned: Content that has been technically manipulated or doctored in a way that misleads users (beyond decontextualised clips), e.g., to falsely suggest the death of a government official or fabricate events where there is a serious risk of egregious harm.	Potential to mislead and risk of egregious harm	None	Human and automated moderation.	External evaluators, community reporting, priority flaggers. Fact-checkers (limited to some countries)

Platform	Definition of AI-manipulated or generated content and mention to AI	Distinction between allowed and banned AI-manipulated or generated content	Rationale for removing AI-manipulated or generated content	AI-related resources	Internal actors	External collaborators
TikTok	Synthetic and manipulated media: "content created or modified by AI technology."	Banned synthetic media... ... showing realistic scenes that are not disclosed or labelled. ... containing the likeness (visual or audio) of a real person, including: (1) a young person, (2) an adult private figure, and (3) an adult public figure when used for political or commercial endorsements, or if it violates any other policy. ...that has been edited, spliced, or combined (such as video and audio) in a way that may mislead a person about real-world events. ... violating other policies (hate speech, sexual exploitation, harassment,...) Allowed synthetic media: Synthetic media showing a public figure in certain contexts, including artistic and educational content.	Integrity and authenticity - risk of harm, abuse or mislead.	None	European Safety Advisory Council; Automated and human moderation.	Safety partners (i.e., fact-checkers)
X (previously Twitter)	Synthetic and manipulated media (as part of misleading media), minimal mention of AI.	Banned media: ... significantly and deceptively altered, manipulated, or fabricated, or ... shared in a deceptive manner or with false context, and ... likely to result in widespread confusion on public issues, impact public safety, or cause serious harm. Allowed: Memes, satire; animations, illustrations, and cartoons; commentary, reviews, opinions and/or reactions and counter-speech.	High-severity violations of the policy; potential to mislead and serious risk of harm.	None	Combination of human and automated moderation.	Partnerships with global third-party experts. Volunteer content moderators via Community Notes (previously Birdwatch), Moments, and Misleading Info Reporting Flow, but limited to specific countries.

TYPES OF ACTIONS

Platform	1. Labelling of AI-manipulated or generated content	2. Downranking AI-manipulated or generated content	3. Demonetisation of AI-manipulated or generated content	4. Strike policy	5. Removal of AI-manipulated or generated content
Facebook	Informational labels "altered content" for manipulated media non-eligible for removal but considered false or partly false by a third-party fact-checker. Informational labels for generated content with Meta AI.	Visibility/distribution in the news feed will be reduced for manipulated media non-eligible for removal, but considered false or partly false by a third-party fact-checker.	Content debunked by fact-checkers is prohibited by Meta's Advertising Standards and Partner monetisation policies. Community Standards compliance is required to monetise content.	Meta's strike policy for violating Community Standards applies. On Facebook, strikes will lead to different restrictions up to disabling accounts.	Manipulated media removal will apply when: 1. It has been edited or synthesised – beyond adjustments for clarity or quality – in ways that are not apparent to an average person and would likely mislead users into thinking that someone said words they did not actually say. 2. It is the product of AI or machine learning that merges, replaces, or superimposes content onto a video, making it appear authentic. 3. It violates Community Standards.
Instagram	Informational labels ("altered content") for content non-eligible for removal. Based on fact-checker ratings. Informational labels for content generated with Meta AI.	Generic approach to misinformation: reducing the distribution of false information (based on fact-checkers' decisions).	Content rated false by a third-party fact-checker is ineligible to monetise. Advertisers must follow Instagram Community Standards.	Meta's strike policy for violating Community Standards applies. Accounts that do not follow the Community Guidelines may be disabled.	Not specific for AI (general approach): content removal will be applied when it violates the Terms of Use, Instagram policies (including Instagram Community Guidelines), or if it is required by law.
YouTube	General approach: Rating system based on fact-checks. The required disclosure of "synthetic content" in Google's political ads in some regions also will apply to YouTube from mid-November 2023 (but not a general policy of the video platform).	Not specific for AI (general approach): removing borderline content (based on external evaluators) from recommendations, reducing its visibility.	Restrictions to monetise AI-generated content. "Programmatically generated" content can violate the repetitious content section on the AdSense guidelines. Monetisation guidelines and ads rules require compliance with misinformation policies, among others.	Strike policy for violating Community Guidelines or copyright violations (up to account or channel termination).	Not specific for AI (general approach): content that violates YouTube policies will be removed.

Platform	1. Labelling of AI-manipulated or generated content	2. Downranking AI-manipulated or generated content	3. Demonetisation of AI-manipulated or generated content	4. Strike policy	5. Removal of AI-manipulated or generated content
TikTok	Synthetic media must be clearly disclosed by the user. This can be done through the use of a sticker or caption, such as 'synthetic', 'fake', 'not real', or 'altered'. TikTok has incorporated a new tool for users to tag AI-generated content, according to press reports. Also labelling by fact-checking partners: prompts to help people reconsider before sharing.	Inconclusive fact-checks and labelled content can become ineligible for recommendation into anyone's 'For You feed' (general approach).	Content is ineligible to monetise if it does not abide the Community Guidelines (general approach).	Strike policy for violating the Community Guidelines.	Synthetic media showing realistic scenes that are not prominently disclosed or labelled in the video. Synthetic media that contains the likeness (visual or audio) of a real person, including: (1) a young person, (2) an adult private figure, and (3) an adult public figure when used for political or commercial endorsements, or if it violates any other policy. Material that has been edited, spliced, or combined (such as video and audio) in a way that may mislead a person about real-world events.
X	Some manipulated media violating the policy will receive a label and/or a warning message instead of being removed. X will provide a link with explanations/ clarifications.	X can reduce the visibility or prevent the content being recommended, turn off likes, replies, and retweets for some manipulated media violating the policy but that was not removed.	Creators' monetisation standards and ads should comply with X Rules.	Strike policy for accounts that have advanced or continuously shared harmful misleading narratives that violate the synthetic and manipulated media policy.	For high-severity policy violation, including misleading media that have a serious risk of harm to individuals or communities, X will require the user to remove this content.

TYPE OF CONTENT

Platform	Organic Content	Advertisement Content
Facebook	Policies against manipulated media apply.	Advertisers placing ads must follow Community Standards and Advertising Standards. Meta prohibits ads that include content debunked by third-party fact-checkers.
Instagram	Policies regarding Terms of Use and Instagram Community Guidelines apply, but nothing specific related to AI.	Advertisers on Instagram must follow Instagram Community Guidelines.
YouTube	Policies against misinformation applies, including manipulated content but nothing specific related to AI. Restrictions to monetise AI-generated content based on the AdSense guidelines. "Programmatically created" or "computer generated" content can violate the repetitious content section.	Ads on YouTube have to comply with Google Ads policies. This requires compliance with misinformation policies (and others). Google's updates to political content policy (that presumably will apply to YouTube since mid-November 2023) force to label synthetic content with misleading potential in political ads, but with a restricted application. Restrictions to monetise AI-generated content based on the AdSense guidelines. "Programmatically created" or "computer generated" content can violate the repetitious content section.
TikTok	The Community Guidelines prohibit synthetic and manipulated media that are not clearly disclosed and violate the previously mentioned rules.	Advertising policies prohibit misleading, Inauthentic, and deceptive behaviours.
X	Synthetic and manipulated media policy + misinformation policy applies. Creators monetisation standards include complying with X rules.	Advertisers must follow X's Terms of Service, X Rules, and all the policies on our Help Center governing use of our services. A tweet that violates rules will be excluded from having ads adjacent to it. Creators monetisation standards include complying with X rules.

CONCLUDING REMARKS

This final section offers the opportunity to express some considerations from compiling this factsheet.

- **Lack of transparency**

  Firstly, we would like to highlight the lack of clarity or transparency surrounding many of these policies. This includes for example the collaboration with experts or fact-checkers, whose scope or nowadays status is not entirely clear in some cases. In another example, the line between banned content and content to be removed is not always explicitly defined. While some platforms specify the content to be banned, it is not always clear whether this content will be removed, tagged or downranked.

- **Issues with the user friendliness of policies**

  As mentioned in some of our previous platform policy papers, navigating platforms' policy pages can often be challenging. This is particularly true for Meta's, where there is some confusion on whether pages apply to Facebook alone or Facebook and Instagram together. Furthermore, clear dates of the various publications are often missing on the platform's policies page. This lack of date leaves users, researchers or any other interested stakeholders uninformed about the most recent measures in place or whether a new webpage has been created instead of updating an existing one.

- **Improved alignment**

  On a positive note, Facebook and Instagram, being both Meta products, have aligned their content moderation policies. Therefore, content that is rated as false or partly false on Facebook will be automatically labelled as such on Instagram, and vice versa. This sort of cross-platform policy harmonisation is highly desirable. In this sense, a positive trend is that these VLOPs cooperate with fact-checkers from the International Fact-Checking Network.

- **Protection of minors**

  It is fair to mention that, at least on paper, platforms do seem to pay special attention to the protection of minors, e.g., preventing synthetic media containing the likeness of a young person (TikTok's case).

- **Lack of specific provisions for AI-manipulated or generated content**

  In another common note with our other studies on platforms' policies on misinformation, the lack of specific provisions for dealing with AI-manipulated or generated content necessitates the application of the general misinformation policy, such as on Instagram and YouTube. In the case of YouTube, this content is indeed addressed, but with a very limited scope: programmatically generated content is mentioned in the AdSense guidelines and synthetic content in Google political ads policy (in force from mid-November 2023), but not in other policies. Other shortcomings of scope include not mentioning AI-generated text or focusing on manipulated rather than generated content. Platforms usually talk about 'manipulated' content, but much of it is newly generated rather than doctored, and therefore needs to be addressed in a different way.

- **Challenges to address AI-generated content**

  Several arguments seem to suggest that AI-generated content is under-regulated by the platforms analysed. Platforms that do mention this emerging technology speak of synthetic content or manipulated "media" referring to pictures or video, but sometimes overlook AI-generated text. Moreover, they do not distinguish AI-manipulated (modified) content and AI- generated content. In short, most policies fail to reflect the new possibilities that generative AI introduces. It is also worth noting that one of the biggest challenges is the detection of AI-manipulated and generated content. When content is difficult to detect, it can hardly be moderated.

- **Subjective premise for moderation of AI-manipulated or generated content**

  When addressing AI manipulated or generated content in their policies, platforms mention as a rationale for moderating content the risk for end-users to be misled. For instance, the danger that the content's recipient doesn't realise that the media has been manipulated or fabricated). Basing content moderation on such a subjective premise can nevertheless be up for interpretation and could be potentially exploited to avoid regulation. For instance, with uploaders alleging that the content is satire or parody, that is permitted on the platforms.

- **Limited updates on disinformation policies**

  On another note, just three platforms updated their counter-misinformation policies this year: X, Facebook, and TikTok. Although only TikTok took the opportunity to further

address more in-depth the generative AI considerations. Google (and YouTube) also recently addressed the issue but with a limited scope: only for political ads. Facebook, for example, updated its misinformation policy last July, but nothing changed regarding AI-manipulated content from its 2020 policy. We wonder whether this is an intentional choice or only a matter of time before the latest developments of AI are reflected in their policies. Whatever the reason for this, considerations regarding the platforms' capacity to make effective policy changes that adapt to new needs inevitably emerge.

- **Focus on labelling AI-generated content**

  After the European Union requested the signatories of the Code of Practice on Online Disinformation to label AI-generated content last June, some platforms took action. For instance, TikTok has already incorporated a new tool for users to tag AI-generated content, furthering compliance with its latest policy update. According to reports, Meta is also working on labels, but just for content generated with their own AI products. In the case of X, some people have seen the Community Notes boost as new possibilities for tagging this type of content, although there wasn't a policy change to date. As for YouTube, its parent company Google announced the obligation to label synthetic content in political advertisements in certain regions. In other words, with a very restricted application that never constitutes a general policy for the entire video platform. While it is too early to assess this approach's effectiveness, we believe labelling should complement other moderation measures. Besides, this measure opens up many questions i.e., to what extent will labelling AI-generated content prevent other, harsher punishments?

- **Action needed before AI Act comes into force**

  On the legislative field, the AI Act will bring new rules and obligations on a risk based approach. The initial proposal foresaw some transparency requirements for deepfakes and AI based chatbot (Art. 52 of the proposal). Since then, the EP has suggested amendments to regulate foundation models including generative AI systems. The amendments include transparency obligations and responsible design and development. The text is being currently negotiated but it will take some time until the adopted legislation comes into force.[4] All the platforms studied in this document are VLOPs and are bound by the DSA. While the DSA is technology-neutral, the power of generative AI brings with it new challenges that may not be totally covered by the adopted rules.

  If platforms do not put their own policies in place, bad actors could take advantage of existing loopholes instead.

RECOMMENDATIONS

- For all these reasons, platforms should update their policies and respond with effective policy changes to meet new needs in the face of rapidly evolving technologies. AI in general, and AI-generated disinformation in particular, poses a general concern, but until now it has only generated low responses. While the COVID-19 pandemic and the resulting infodemic led to the development of new content moderation policies in the health domain, a similar proactive approach has not been the case for the emergence of a technology as disruptive as generative AI.
- In this context, platforms should also enhance cooperation with external collaborators and experts in AI, following the collaboration model implemented during the COVID-19 pandemic, with medical experts to combat infodemic. Besides, they should encourage the creation of information AI internal resources such as those that Facebook put in place for COVID-19 or climate change mis- and disinformation.
- In the DSA, risk assessment will be one of the main instruments to fight disinformation. The development of a framework on how to apply this assessment specifically to AI-generated content would be desirable and helpful, to provide guidance and prevent arbitrariness in the assessment process.
- On a final note, AI-generated content also brings new challenges to regulate end-user's role on these platforms. For example, TikTok, underlines user's responsibility by labelling AI content obligations, aligning with what the European AI Act foresees. In the meantime, the European Union is leaning on signatories to its Code of Practice on Online Disinformation to label deepfakes and other AI-generated content, which would include the platforms that abided by the code. The burden of responsibility between all AI manipulated or generated content relevant stakeholders (including users) needs to be more strongly and clearly regulated.

4 Currently, the European Commission, the Council and Parliament are negotiating to produce an agreed version of the text, in the so called trilogue. It remains to be seen what will come out in the final text.