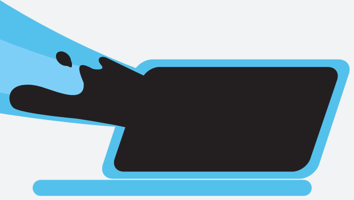
June 2023

PLATFORMS' POLICIES ON HEALTH MISINFORMATION



EU DisinfoLab

TABLE OF CONTENTS

Author and affiliation: **Maria Giovanna Sessa** at EU DisinfoLab

A special thank you goes to reviewers: **Amaury Lesplingart** at CheckFirst and to **Carlos Hernández-Echevarría** at Maldita

INTRODUCTION

- The present factsheet delves into platforms' policies on health misinformation, focusing on Facebook, Instagram, YouTube, TikTok, and Twitter. Very Large Online Platforms (VLOPs) have responded to the COVID-19 pandemic and related infodemic with specific legislation, making the topic of medical and health-related misinformation one of the most advanced topic-specific policies.
- Currently, the sense of urgency of the pandemic has faded, but the frameworks developed can be applied to other emergencies and the risk of imminent harm, which remains the fundamental ratio to recognise and address misinformation in general, and health misinformation specifically. Overall, vaccine-related misinformation continues to abound, making it crucial to maintain and enforce a solid policy that protects individuals and communities.
- A necessary disclaimer is that it is not always possible to distinguish between general misinformation policy and health-specific considerations. This document tries to skim through the various elements of the platforms' misinformation policies to focus exclusively on relevant health- and medical misinformation elements. Therefore, in a long list of prohibited items, it will focus only on those that would have an impact on health, thus mentioning, for example, diet supplements, but not weapons.
- The next pages want to offer a cheat sheet to navigate how five platforms define and address health misinformation, which actions are put in place to limit the impact of health misinformation, and content is allowed to be published and advertised in this regard. Moreover, for each category analysed, it tries to highlight the common traits across platforms.

DEFINITIONS AND ACTORS

Platform	Definition of health misinformation	Rationale for removing health misinformation	COVID-19 misinformation resources	Internal actors	External collaborators
Facebook	Content determined false by an authoritative third party.	Risk of imminent physical harm.	Meta's COVID-19 Information Centre.	Oversight Board; human content moderators.	Health authorities; third-party fact-checkers.
Instagram	Content determined false by public health authorities during public health emergencies.	Risk of imminent physical harm, including risk of getting/spreading a harmful disease or refusing a vaccine.	Meta's COVID-19 Information Centre.	Oversight Board; human content moderators.	Health authorities; third-party fact-checkers.
YouTube	Certain types of misleading or deceptive content.	Serious risk of egregious harm.	COVID-19 Medical Misinformation Policy.	Human content moderators.	Health organisations; third-party fact-checkers.
TikTok	False and misleading content.	Significant harm to individuals or community, including serious physical injury, illness, death, severe psychological trauma, and public distrust in scientific bodies.	COVID-19 page in the Safety Centre.	European Safety Advisory Council; human content moderators.	WHO and Team Halo; third-party fact-checkers.
Twitter	Misleading content, content determined false by external subject-matter experts.	Causing serious harm and impacting public safety, serious harm during public health emergencies.	No longer enforcing COVID-19 misleading information policy since 11 November 2022. The crisis misinformation policy (May 2022) still mentions public health emergencies.	Volunteer content moderators via Community Notes (previously Birdwatch), Twitter Moments; Misleading Info Reporting (unavailable for EU countries); Human content moderators.	External, subject-matter, third-party experts.

TYPES OF ACTIONS

Platform	1. Labelling of health misinformation	2. Downranking of health misinformation	3. Demonetisation of health misinformation	4. Strike policy	5. Removal of health misinformation
Facebook	Informational labels; third-party fact-checker rating system.	Restrictions include reduced distribution and removal from recommendations.	Disproven medical claims, including anti-vaccination claims.	Meta's strike policy for violating Community Standards.	Harmful health misinformation about vaccines, miracle cures, and during public health emergencies.
Instagram	Informational labels; third-party fact-checker rating system.	Restrictions include reduced distribution and removal from recommendations.	Disproven medical claims, including anti-vaccination claims.	Meta's strike policy for violating Community Standards.	Content interfering with COVID-19 vaccine administration, hate speech related to COVID-19, outing individuals for having COVID-19.
YouTube	No "judgement on the accuracy of any video"; no third-party fact-checker rating system.	Removing borderline content from recommendations.	Withholding, limiting, or suspending channel revenue for violations of guidelines prohibiting medical misinformation.	Strike policy for violating Community Standards.	Misinformation promoting dangerous remedies, contradicting expert consensus or health authorities' guidance.
TikTok	Warning labels by third-party fact-checking partners; banners on videos and reminders on searches for COVID-19-related content.	Limiting distribution of inconclusive content in "For You" feed.	Channel suspension for violation of TikTok Advertising Guidelines or other standards, including pandemic, vaccine, and medical misinformation.	Strike policy for violating the Community Guidelines.	Medical misinformation causing harm to physical health.
Twitter	Labelling content (unclear if applicable to medical misinformation).	Reduced visibility for labelled tweets (unclear if applicable to medical misinformation).	Suspension, shadow-banning, or removal for misleading claims with potential to cause harm, including miracle cures.	No specific strike policy for medical misinformation.	Risk of immediate and severe offline consequences (medical misinformation not mentioned).

TYPE OF CONTENT

Platform	Organic Content	Advertisement Content
Facebook	Detailed COVID-19 and Vaccine Policy, prohibiting misinformation related to the transmission and immunity, cures and prevention methods, discouraging good health practices, and false health information (especially about vaccines).	The Advertising Standards consider unacceptable content (misinformation, vaccine discouragement, discriminatory practices based on disability, medical, or genetic condition, and inflammatory content based on disability or serious disease); deceptive content (health-related unrealistic outcomes); dangerous substances; and objectionable content (health-related personal attributes and appearance, and commercial exploitation of crises).
Instagram	Detailed COVID-19 and Vaccine Policy, prohibiting misinformation related to the transmission and immunity, cures and prevention methods, discouraging good health practices, and false health information (especially about vaccines).	The Advertising Standards consider unacceptable content (misinformation, vaccine discouragement, discriminatory practices based on disability, medical, or genetic condition, and inflammatory content based on disability or serious disease); deceptive content (health-related unrealistic outcomes); dangerous substances; and objectionable content (health-related personal attributes and appearance, and commercial exploitation of crises).
YouTube	Detailed COVID-19 medical misinformation policy, prohibiting misinformation related to the treatment, prevention, diagnostics, transmission, and denial of COVID-19's existence.	YouTube ads have to comply with Google Ads Policies, prohibiting dangerous products and services, inappropriate content, health-related service misrepresentation, and restricting healthcare and medicines.
TikTok	The Community Guidelines prohibit harmful medical misinformation and health-related hateful behaviour.	Advertising Policies prohibit health-related discriminatory content, misleading claims, inappropriate content, data collection, and restrict weight control/management and body image, and COVID-19- and vaccine-related content.
Twitter	No reference to health-related content (except for a prohibition to attack others based on disability or serious disease).	Unclear if medical misinformation applies to inappropriate content. The healthcare policy includes medical products claiming to diagnose, cure, treat, or prevent diseases, diet products, healthcare and wellness substances, and medical and cosmetic services.

CROSS-PLATFORM COMPARISON

Common traits	Facebook	Instagram	YouTube	TikTok	Twitter
Definition of health misinformation based on falsity and/or mislead	X	X	X	X	X
Definition of health misinformation based on third-party assessment	X	X			X
Rationale for removing health misinformation based on risk of harm	X	X	X	X	X
Specific COVID-19 resources	X	X	X	X	
Human content moderators	X	X	X	X	X
Collaboration with health authorities and organisations	X	X	X	X	*
Collaboration with fact-checkers	X	X		X	*
Labelling of health misinformation	X	X	**	X	X
Downranking of health misinformation	X	X	X	X	X
Demonetisation of health misinformation	X	X	X	X	X
Strike policy	X	X	X	X	
Removal of health misinformation	X	X	X	X	
Prohibition of health misinformation	X	X	X	X	***
Advertising standards for health misinformation	X	X	X	X	X

*Unclear whether Twitter relies on health authorities and organisations, and fact-checkers for (health-related) content moderation.
**ClaimReview panels are (extremely rare and) associated with searches and not videos.
***Twitter removes content that violates its rules, but medical misinformation is not explicitly mentioned.

CONCLUDING REMARKS

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

- As the underlying ratio for health misinformation content moderation policies is the risk of harm, we notice that these policies developed greatly in the context of the COVID-19 pandemic, as well as their overlap with policy around illegal practices such as hate speech and discrimination.
- It is often difficult to navigate the platforms' policy pages – especially Meta's, where there is some confusion on whether pages apply to Facebook alone or Facebook and Instagram together. Overall, there is a lack of linearity regarding "what resource does what", i.e., the "about.meta.com", "transparency.fb.com", or the "oversightboard.com" sites. Furthermore, clear dates of the various publications are often missing, so that one is unaware what is the latest measure in place, or if a new webpage has been created rather than updating an existing one.
- On a more positive note, Facebook and Instagram, being both Meta products, aligned their content moderation policies. Therefore, identical content that is rated partly false on Facebook will be automatically labelled as such on Instagram too, and vice versa. This sort of cross-platform policy harmonisation is highly desirable. In this sense, a positive tendency is that these VLOPs cooperate with fact-checkers from the International Fact-Checking Network.
- It is extremely concerning that Twitter does not have a COVID-19 and health misinformation policy anymore. In general, the platform seems to use terms such as "misinformation", "misleading content", or "false content" interchangeably. Moreover, the platform's lack of fact-checking by professionals is alarming, and the bottom-up approach to content verification envisioned by Community Notes has already revealed to be a failure, highjacked by disinformation spreaders and believers. Another contradiction is that Twitter's crisis misinformation policy but does not address content about COVID-19. Moreover, Twitter does not enforce the crisis misinformation policy on "personal anecdotes or first-person accounts". This could clearly generate a loophole if, for instance, someone was to blame a vaccine for alleged symptoms.
- In spite of efforts to counter misinformation in general, and health misinformation in particular, platforms often adopt an excessively mild approach, which seems more an attempt not to alienate some users rather than protecting the whole community. For instance, YouTube states that "one person's misinfo is often another person's deeply held belief, including perspectives that are provocative, potentially offensive, or even in some cases, include information that may not pass a fact checker's scrutiny". This is a highly problematic sentence, which contributes to the information disorder by blurring the differences between fact and opinion.
- On a final note, it is fair mentioning that, at least on paper, platforms do seem to pay special attention to the protection of minors, e.g., preventing harmful behaviour linked to body image, eating disorders, or self-harm. This is especially true for TikTok, whose primary users are very young.