fake news

how platforms face disinformation

intervozes
coletivo brasil de comunicação social

RESEARCH
FAKE NEWS: HOW PLATFORMS FACE DISINFORMATION

AUTHORS
Bia Barbosa, Helena Martins e Jonas Valente

GRAPHIC DESIGN AND DIAGRAMMING
Oficina Sal

INTERVOZES - BRAZIL SOCIAL COMMUNICATION COLLECTIVE
BOARD OF DIRECTORS
Alex Hercog, André Pasti, Bruno Marinoni, Marcos Urupá,
Mônica Mourão, Paulo Victor Melo e Tâmara Terso

EXECUTIVE COORDINATION
Gyssele Mendes, Iara Moura, Maria Mello, Marina Pita, Olívia Bandeira e Pedro Ekman

SUPPORTED BY (FUNDING ORGANIZATION)

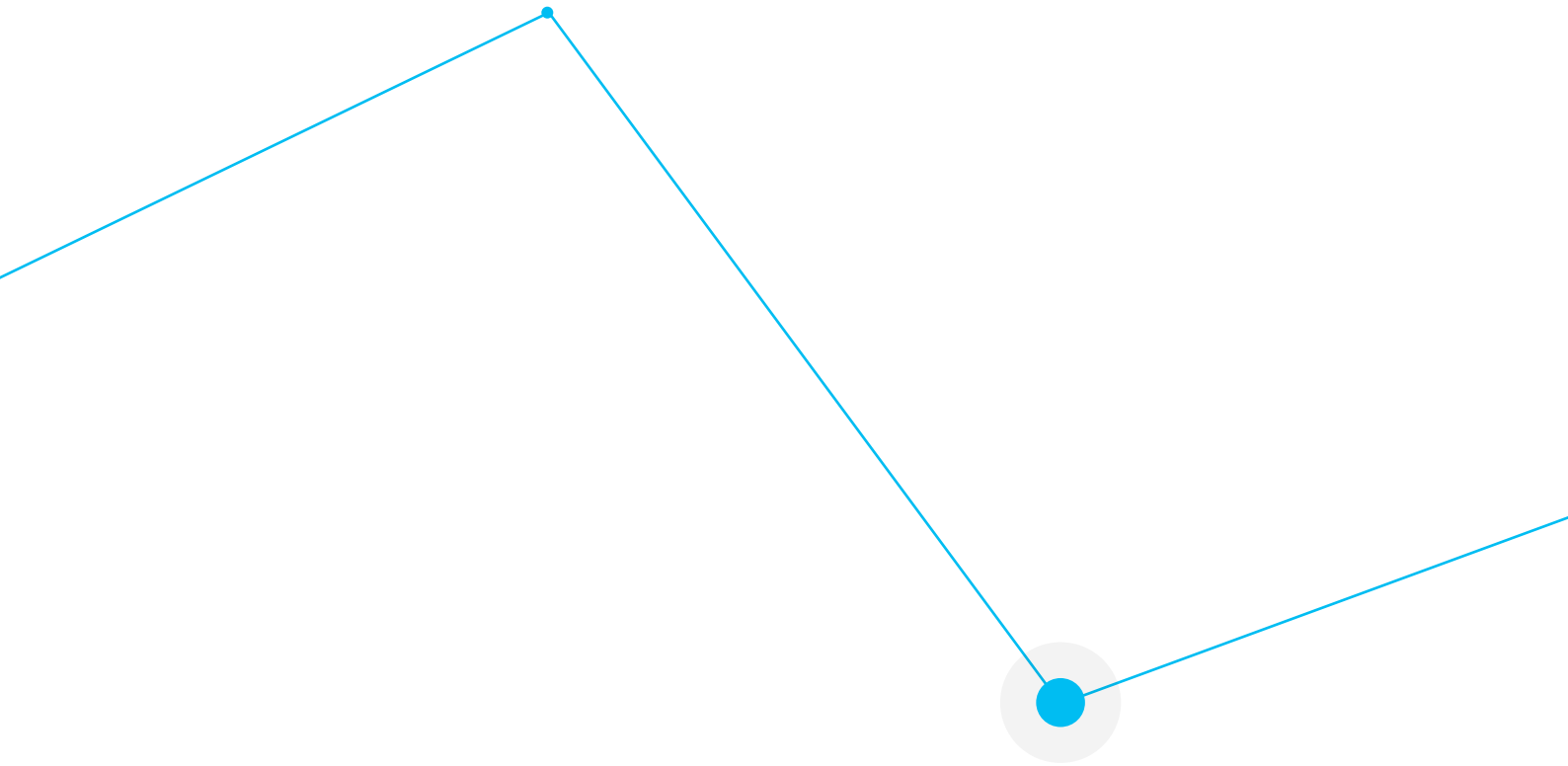
FORD FOUNDATION

intervozes.org.br | intervozes@intervozes.org.br

facebook.com/intervozes
instagram.com/intervozes
twitter.com/intervozes
youtube.com/intervozes

fake news

how platforms face disinformation



executive summary_

1. Summary of the research

The research analyzes disinformation as a strategy to obtain political and economic gains.

It relates this phenomenon to the current configuration of the Internet, which is characterized by the participation of digital monopolies that act with the aim of capturing users' attention in order to collect and treat personal data which are later used for the creation of profiles and for directing messages, such as advertisements or political propaganda.

It considers that the platforms business model favors the occurrence of disinformation that, although not introduced by them, as it can be seen also in the history of traditional media such as radio and TV, becomes more constant, comprehensive, penetrating and of rapid circulation due to the forms of production, circulation, algorithmic mediation, and access to information that are typical of digital platforms.

In consideration of this relevance, we have analyzed the main social networking platforms in operation in Brazil: Facebook, Instagram, WhatsApp, YouTube, and Twitter.

The measures adopted by each of these companies to combat disinformation since 2018, when the phenomenon was already considered a worldwide problem, until the end of 2020 — thus also including actions that are justified by the context of the Covid-19 pandemic — are described.

For description and analysis of such measures, four categories were defined: 1. approach to the phenomenon; 2. Content moderation; 3. promotion of information and transparency; and 4. related measures.

The measures are analyzed in view of the context and references expressed in existing rules of platforms on disinformation, documents developed by international authorities and human rights experts, as well as proposals already prepared by Intervozes itself.

The tables below present the main measures adopted by digital platforms in this field, including related measures which indirectly affect disinformation.

Plataforma	Approach to the phenomenon	Content moderation
Facebook	It does not have a specific policy, nor does it work with a definition of disinformation of its own. It presents the strategies adopted against the phenomenon in a summarized version in the Community Standards. The elaboration of measures to be adopted on the subject is carried out by the "Global Content Policies" team. In 2020, it created the Content Oversight Board, with members from outside the company, who are responsible for in-depth review of cases (not only about disinformation).	It admits complaints and uses automated systems to analyze posts. It submits content for verification by partner fact-checking agencies and, in case of being considered as "false" or "partially false", the content receives a specific label and starts to be presented with "related articles" produced by checkers. Statements and postings by political leaders are exempted and do not undergo verification processes by partner fact-checking agencies. Videos and images can be labeled as "manipulated", "taken out of context" or "false". There is no removal from the platform, but the circulation of false postings is reduced in feeds and recommendation mechanisms. Cases of disinformation that cause violence or harm, bring false information about vaccines, or jeopardize electoral processes are removed. Advertisements with false messages may not be published and reoffending advertisers may be sanctioned.
Instagram	It follows the same general approach to the phenomenon adopted by Facebook, which is its controller. The Terms of Use determine that the user may not "do anything illicit, deceptive, fraudulent or for an illegal or unauthorized purpose. It has no institutional structure or specific processes to deal with misinformation. They apply the same processes and procedures adopted by Facebook, including the content review teams and the Supervisory Board.	It carries out content analysis processes from fact-checking agencies, which classify publications into different categories. Contents are flagged when they are verified as disinformative and may have their circulation reduced (less exposure in the feed and stories, removal from the "explore" and hashtags pages). Contents related to false treatments on Covid-19, conspiracy theories or false claims registered as harmful by health authorities are removed, as well as ads and hashtags that might promote disinformation on the subject. In searches related to the virus, it prioritizes information from health authorities. During elections, if a content is classified as false, a gray filter is displayed over the image and alerts are shown to the user before the user shares the publication. Transparency tools for ads have been adopted in several countries.
WhatsApp	It presents itself as a platform of encrypted private messages that, as such, does not access, moderate, judge, verify, block, or remove any content, which is contradictory to the existence of groups and transmission lists that end up making it function as a social network. It does not acknowledge the frequent use of the platform for disinformation. In accordance with such understanding, it does not have a defined policy, processes and/or institutional structure nor a concept of disinformation.	It officially informs that it does not moderate content. For example, it does not flag content as disinformative. Regarding moderation, we only point out that, since 2018, it has been acting to reduce circulation, by establishing limits for the forwarding of messages. First, the limit of forwarding messages in one go was set at 20 chats. Then, 5. Since the pandemic, highly forwarded messages can only be forwarded to one contact at a time. This policy, however, is usually not associated with combating disinformation. Such connection was only made clear in the context of the pandemic.

Promotion of information and transparency	Related measures
It offers the user an icon to search for more information about the source of the post. In the context of the Covid-19 pandemic, it provides information from health authorities in the search results (directing users to the WHO website) and has launched the Coronavirus Information Center, available at the top of the News feed. People who have interacted with "harmful fake" content receive messages. It sponsors media education initiatives, has developed materials and campaigns on the topic, and has partnered with fact-checking agencies to provide tips on content verification. It supports journalistic experiences aimed at expanding the circulation of professional news on the platform. It provides data to 60 researchers around the world.	It has adopted a policy of banning hate speech in 2018 and improved automated systems to identify violent videos, restricting certain live broadcasts. Searches related to white supremacy and exaltation of Nazism receive indications of materials to combat hatred. In 2020 it began to remove holocaust denial content. In the same year, the QAnon group, which disseminates conspiracy theories, was also banned. The platform also removes false accounts which spread spam or are considered to be "coordinated non-authentic behavior". It tries to verify the authenticity of the accounts at the moment of their opening and maps signs of malicious conduct on newly opened accounts. It also reduces the circulation of accounts that spread low quality content (click-hunters or the ones that takes to sites with malicious and shocking ads). It started to restrict access to data collection by third-party apps. As a measure to protect electoral integrity, it provides a political ad archive with information on the identity of advertisers, payment, and scope of content, in addition to labeling these pieces as "electoral propaganda". Users can tell whether they want to receive this type of advertisement
Computerized systems search for "image matching" to label content already classified as fake, even when they were initially published on Facebook (and vice versa). It provides counterpoint texts prepared by fact checkers, with "clarifications" or official information on the subject. Vaccine searches direct users to the WHO website. In the context of the Covid-19 pandemic, it does the same and started to make information from the agency available at the top of the users' feed, besides offering images about prevention measures for sharing in the stories.	The Community Guidelines condemn hate speech, unwanted messages sent repeatedly, and attacks on people with intent to embarrass them. False accounts are removed, mostly at the time of the attempt to open them. Accounts whose identity is verified can obtain a verification flag. Accounts or networks of accounts falling into the category of "inauthentic behavior" are removed. Accounts that show certain numbers of violations during a certain period may be deactivated, with the right to appeal. Promotion, offering or trade of false user ratings are prohibited; and inauthentic likes, followers and comments aimed at "boosting popularity" are removed. It has increased measures against bullying and harassment, giving greater control over user interactions and no longer displaying the number of likes on posts.
In the application or in the official blog, one can find references to disinformation, but the information is scattered and not very accessible to a common user, although the platform has the possibility of direct contact with the user, through notifications. Since the 2018 elections, institutional partnerships have been made to verify content and conduct surveys, but their scope and effectiveness are questionable. Likewise, actions to raise awareness among the broad public are limited. Information promotion measures have been expanded in the pandemic, with the creation of message channels. In the same way, for the 2020 elections, an agreement was signed with TSE and a chatbot was created for complaints and stickers on conscious voting. The user must add the account number and make contact. There is a lack of transparency regarding the practices adopted by WhatsApp, such as the prohibition of account creation or the removal of accounts, also regarding the reports on their measures.	Most of WhatsApp's actions fall under "related measures", starting with the definition of uses involving false, incorrect, or misleading statements as violations in its Terms of Service. In addition to the limitation of message forwarding, it has flagged highly forwarded messages with one or two arrows, indicating greater viralization. It has also removed the button that allowed fast content forwarding and inserted a magnifying glass next to the messages marked with two arrows. It uses spam handling tools and advanced machine learning to remove automated mass messages and ban user accounts with inappropriate behavior, such as sending mass messages and creating multiple accounts.

Plataforma	Approach to the phenomenon	Content moderation
YouTube	It does not work with a specific definition of disinformation and does not have a specific policy for the subject. But restrictions to the dissemination of disinformation are found in different policies, such as on harm and manipulated media. Content that violates the Community Guidelines is removed and notified. Recurring cases suffer publishing and monetization restrictions, and the channel can be excluded. It encourages the community to report inappropriate content and internal teams analyze the cases. It recently launched the "Myths and Facts about Disinformation" page, which gathers measures it adopts against this issue.	It removes content that violates its guidelines and reduces the reach of "harmful disinformation" and "borderline content" by no longer recommending them. Edited and tampered videos are prohibited by the policy on "manipulated media", as well as channels that falsify their identity. It removes content and channels that disrespect their politics for elections, including videos that aim to undermine the confidence in the polls. It began to remove misleading information about Covid-19, especially medical disinformation, promotion of dangerous drugs, and information about the origin of the virus. It has adopted a strong policy against anti-vaccine videos. It recommends producers to check facts and use health authorities as a source. Not all advertising around the pandemic has been restricted and videos that follow community standards can receive ads. Decisions on compliance with the guidelines are made based on content analysis regardless of who speaks or publishes such materials, including high authorities. It has a Threat Analysis Group to identify disinformation practices promoted by governments.
Twitter	It does not have a policy for validating the authenticity of content, nor does it work with a definition of disinformation. It focuses on providing context to help people to make up their minds about the content of a contested post, only acting directly in cases of possible damage caused by manipulated media, misleading information about electoral processes, and most recently related to Covid-19. It prohibits the use of network manipulation robots. It has a Reliability and Security Council that also analyzes the issue.	Manipulated media, including deepfakes, are flagged, may have their visibility reduced, receive a link with further explanations or may be removed in case of risk of immediate and serious harm. Content that may mislead people about when, where, and how to vote is also prohibited, as part of the policy of integrity in elections, which in the U.S. marks candidate posts. False information intended to undermine public confidence in the elections may be labeled or removed. Political and state media advertisements are no longer allowed. It monitors trends and spikes in election conversations to detect manipulative activities and remove false accounts. Posts by global leaders, even if they are uninformative, can go on air if they are of clear public interest. The post is displayed with a warning and it is not possible to interact with it. Following the Covid-19 pandemic, it extended the definition of "harm" to include content that is contrary to the recommendations of health authorities and it removes posts with a clear call for risky actions. When it poses a lower risk of harm, the post receives a label or warning. The advertising policy is only authorized for public and private services that have changed with Covid-19.

Promotion of information and transparency

Related measures

It states that priority is given to what it calls “authorized voices”, such as journalism channels, in the search for information about events and politics and in the “upcoming videos” on these topics. If the user searches for journalistic content, it shows a section with “Main news” for him. The “Latest News” section is displayed for important events. It displays information panels with context data from authorized sources on historical, scientifically proven topics and conspiracy theories. Depending on the search, fact checks by independent editors are also displayed. There is no check on each individual video. The Covid-19 pandemic has gained special sections on all these tools. Regarding the U.S. elections, it offers additional data on candidates. In Brazil, it made live videos about disinformation with the TSE. Through the Google News Initiative, it supports digital journalism in combating disinformation and actions of media education.

It has its own policy against hate speech and one against harassment and threats. Supremacist videos are forbidden and those that cannot be characterized as hate but come close to it have restrictions on engagement, recommendation, and monetization. An alert is displayed before it begins playing. In 2020, it began to remove content related to conspiracy theories that result in acts of violence. Any person or channel that intimidates, persecutes, dehumanizes, and encourages violent behavior, also through comments, can be banned from the network. It also has a policy to deal with issues such as privacy and people’s defamation, in order to receive complaints. It does not authorize the distribution of dangerous or harmful content and has a policy against spam, falsification of identity and false involvement/engagement. It restricts access to the service through any automated means (bots).

It has the #KnowledgeFacts tool, which is displayed in searches associated with vaccines and the Covid-19. Non-reliable health information does not appear in search results, which prioritize official sources. It has added a special tab about the pandemic in the #Explore function, with reliable sources, public releases, and health journalists. It uses a label on posts about Covid-19 and 5G, encouraging fact-checking and, in the “Events” resource, it also selects reliable information and makes it available at the top of the timelines. It has increased the number of verified accounts in this respect. It has opened the platform for real-time tracking of posts about the Covid-19 by developers and researchers. It creates special pages about the elections with official and reliable information. It supports independent journalism, fact-checking, and media education initiatives, also providing credits in advertisements.

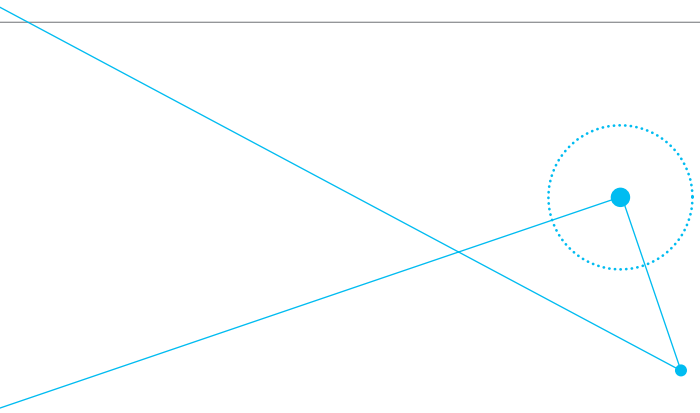
Its policy against the spread of hate bans content with violent threats, which promotes or expresses a desire for death or aggression, the spread of stereotypes, fear, and harassment. Behaviors considered abusive are also not allowed. Violators may have to exclude content, spend time without publishing or interacting, or be totally suspended. It fights spam and malicious automation by monitoring and challenging suspicious account activities, such as through identity verification. The focus is on stopping false accounts, inauthentic engagement and coordinated activities to artificially influence conversations. Suspicious profiles have reduced visibility and audits are performed on existing accounts. It discloses files of coordinated activity posts supported by governments to give transparency to attempts of manipulating the network.

_conclusions

The study concludes that the analyzed digital platforms do not present policy and structured processes on the issue of disinformation nor do they develop punctual and reactive actions to combat the phenomenon. In general, they avoid the analysis of disinformative content, but end up doing so. None of the companies has reported to work with a unified concept on disinformation. Besides mentions to it, the use of various terms, such as false news and misleading information, has been found. In spite of the polysemy of the concept and the intense debate both in society and in academic circles about it, the explicit statement of an understanding would be an advance as a way to give transparency about what kind of content can be impacted by the measures of analysis, flags or sanctions. Not addressing it directly does not mitigate the problem and opens room for others, such as unilateral and unclear decisions. Moreover, none of the companies has reported having a specific structure to address the issue of disinformation, which may hinder the coordination of initiatives within each organization.

Regarding the moderation of disinformative content, there is an increase in automated decisions by the platforms, but it is not clear, even in the explanations given by the corporations for this research, the criteria for the use of natural persons in the analysis and decision on measures for content moderation. The verification of content, mainly by external agencies, is practiced by most platforms. When it occurs, the checking is carried out according to the platform's categorization, which is a necessary guideline to avoid different treatments by each checker. Even so, the complexity of analyzing the "shades of grey" between one extreme and the other naturally gives rise to the risk of questionable assessments, which is why the verification should have effective process mechanisms for the prevention of abuses and errors. On platforms where content is confirmed, it gets flagged. Political and academic players have pointed out the inefficiency of this mechanism; studies have shown that the alerts even draw more attention to the content. Even with this "side effect", we consider it a necessary tool.

Moderation also covers advertisements and promoted content. In these cases, there are restrictions to content already considered disinformative. The extension of the measures on



For most of the period analyzed by the survey, platforms have been resistant to removing disinformative content

disinformation for advertisements is critical. This is where the economic dimension of producing and spreading false news as a business lies, which is used by ''factories of misleading content'' and also by the platforms themselves, that benefit from the charge of promotion. In the case of ads with speeches by politicians, the measures conflict with the exceptions adopted by some companies to these agents' speech.

For most of the period analyzed by the survey, platforms have been resistant to removing disinformative content as they do for other categories, based on what is unilaterally defined in their guidelines. But this has begun to change in 2020, in the context of the new coronavirus pandemic, with the spread of disinformation about the disease and about "miraculous means" of cure. The serious risks to the population's health have put pressure on the platforms to respond more quickly and harshly on this. In this scenario, the acceptance of exceptional situations of withdrawal in cases of evident risk of serious harm seems like a reasonable possibility, provided it is connected to rules of a proper process that allow for contestation, evaluation of resources by people and compensation in case of error in the applied moderation, which is not guaranteed in any of the analyzed platforms.

The reduction of the reach, which is the most frequent measure in cases of misinformation, becomes almost a condition of exclusion and appears to be a measure of high impact on the contents, but it has not been fully proven to be effective. Studies mentioned in the research point to a reduction in the speed of disinformation spreading, but the reach limitation does not prevent such disinformative contents to continue circulating. It is necessary to deepen studies on the subject, with emphasis on research that can effectively focus on the operation of professional groups that use the platforms with systems that enable to avoid these limitations.

Part of the analyzed platforms provides more information in the case of verified contents. One form of such action is the display of articles from fact-checking agencies with information on a given subject. Another type is the offer of "context" information or "reliable sources". Contents of journalistic and official media can also be provided. In the already mentioned pandemic context, several mechanisms of access and recommendation of official information on the subject have been implemented, especially from the World Health Organization and national health authorities.

From the point of view of information on how it deals with disinformative content, the platforms' low transparency is striking. The existing measures are not presented in an organized manner and much of the work of this research was exactly the search and organization of information scattered in news on official sites and on "help" or FAQ sites of each of the companies – the exception was YouTube, which only in October 2020 began to dedicate a page to the subject. The lack of recognition of the seriousness of the problem and the low involvement of the platforms in its combat ends up creating obstacles for effective communication with users on the subject. However, the most problematic element in terms of transparency is perhaps the lack of balance of the reported actions. Once again, information is scarce and scattered, so that the assessment of the concreteness of the actions is restricted to the allegations of the companies, coming closer to promises that cannot be proved. There is also no evaluation of the effectiveness of what has been implemented or proper process, a mechanism that could enable users to be notified and be able to defend themselves in processes of content moderation or in possible sanctions applied by the platforms on their posts or accounts.

We concluded that the actions still require organization, consistency, transparency, and evaluation. The issue of disinformation needs to be effectively recognized, communicated and faced by the platforms, which requires the reviewing of the structure and business model of these companies, otherwise they will continue to offer remedies that are unable to stop a form of communication that today has become one of the main challenges for democracies around the world.

Summary of measures adopted by digital platforms (2018-2020)

		Facebook	Instagram	WhatsApp	YouTube	Twitter
APPROACH TO THE PHENOMENON	Specific and structured policy to fight disinformation					
	Adopted and defined concept of disinformation					
	Mention of disinformation in guidelines or policies			x	x	x
	Related measures focusing on disinformation	x	x	x	x	x
	Specific institutional structure for the topic					
CONTENT MODERATION*	Verification of disinformative contents by own team				x	x
	Verification of disinformative contents by third parties	x	x	x		x
	Removal of disinformative contents				x	x
	Reach reduction	x	x		x	x
	Demonetization and restrictive measures in advertisements and promoted content	x			x	x
	Suspension of accounts due to disinformation				x	x
	Appeal to report of disinformation				x	x
	Notification of users in case of sanction for disinformation	x	x		x	x
	Appeal to the platform in case of sanction for misinformation				x	x

(*) On YouTube and Twitter, moderation occurs when the platform matches certain disinformation with its policy against content that generates harm or affects elections.

		Facebook	Instagram	WhatsApp	YouTube	Twitter
INFORMATION PROMOTION AND TRANSPARENCY	Flagging of disinformative contents	x	x			x
	Availability of information on content verification	x	x	x		x
	Availability of context content or official sources	x	x	x	x	x
	Advertisement transparency measures	x	x		x	x
	Information channels on measures against disinformation					
	Statement on measures taken against disinformation in transparency reports					
	Media education initiatives	x	x	x	x	x
	Support projects for professional journalism	x	x	x	x	x
RELATED MEASURES	Ban of inauthentic accounts	x	x		x	x
	Restrictions to accounts and automated tools			x	x	x
	Preventing the propagation of hate speech	x	x		x	x
	Restriction of manipulated media	x	x		x	x
	Restriction to messages circulation			x