A GUIDE TO HATE SPEECH

FGV DIREITO SP
CENTRO DE ENSINO
E PESQUISA EM INOVAÇÃO

CONIB
Confederação Israelita do Brasil

SUMMARY

What are hate speeches? 4

What makes hate speech severe? 6

What is the Structured Set of Variables? 8

Identification 9
I. Target 9
II. Message 10
III. Intentional Context 10

Evaluation 11
IV. Situational Context 12
V. Speaker 12
VI. Audience 12
VII. Message's vehicle 13
VIII. Socio-Historical Context 13
IX. Consequences 13

Sanctions and Regulation 14
X. Prevention policies 14
XI. Counter speech 15
XII. Removal 15
XIII. Censorship 15
XIV. Civil Sanctions 16
XV. Criminal Sanctions 16
XVI. Administrative Sanctions 16
XVII. Private Sanctions 17

Infographics of the Structured Set of Variables 18

Work Group 19

Opinions expressed in this Guide to Hate Speech are exclusively those of the authors and do not necessarily reflect the institutional views of FGV.

The growth, aggravation and complexity of cases involving hate speech, especially in social networks, has reinforced the need for a set of conceptual tools that may help corporations, NGO's, the civil society and state institutions deal with, mitigate and solve such cases. In this regard, and extremely concerned with this very relevant matter, CONIB has established a partnership with the Getulio Vargas Foundation São Paulo Law School (FGV DIREITO SP) for a project which aims to clarify the concept of hate speech by building a set of variables to identify, evaluate and sanction it.

This Guide is a result of a research, conducted from 2017 through 2019, by the Center for Education and Research on Innovation of the FGV DIREITO SP **(CEPI/FGV).**

The aim of the research was to elucidate the **legal concept** of hate speech, by building a Structured Set of Variables (SSV) for the identification, evaluation, regulation, and sanctioning of this kind of speech.

The SSV was based on the study of several judicial precedents and lawsuits, statutes, as well as academic literature, both from Brazilian and international sources. The work should not be regarded as conclusive. Rather, it should be taken as a landmark for ongoing discussions on hate speech, based on a set of organized topics, **a practical tool for lawyers, police officers, judges and prosecutors,** and a **research agenda**.

Since the SSV incorporates countless sources and deals with a highly complex subject, its application depends on a theoretical stand regarding some of the concepts adopted, e.g., whether certain groups are vulnerable, whether certain paradigmatic types of messages are hate speech, whether some situational contexts allow for certain manifestations to be tolerated, or whether sanctions such as removal and censorship are acceptable.

The aim of CEPI/FGV is not to take a stand on all issues related to the subject; it is rather to organize the debate in a systematic way after providing basic conceptual clarification, by offering a preliminary definition, as well as explaining its "umbrella character" and its connection to the harm endured by members of vulnerable groups.

This Guide was organized as a simplified summary of the Structured Set of Variables. Therefore, it does not provide a conclusive test nor offers easy and definitive answers. It is, above all, a guide for discussion and reflection.

What are **hate speeches?**

Hate speeches are forms of expression that **negatively evaluate** a **vulnerable group** or an **individual who belongs to that group**, in order to establish that such group and its members are **less deserving** of rights, opportunities or resources than other groups or members of other groups and, hence, legitimate the practice of **discrimination** or **violence**.

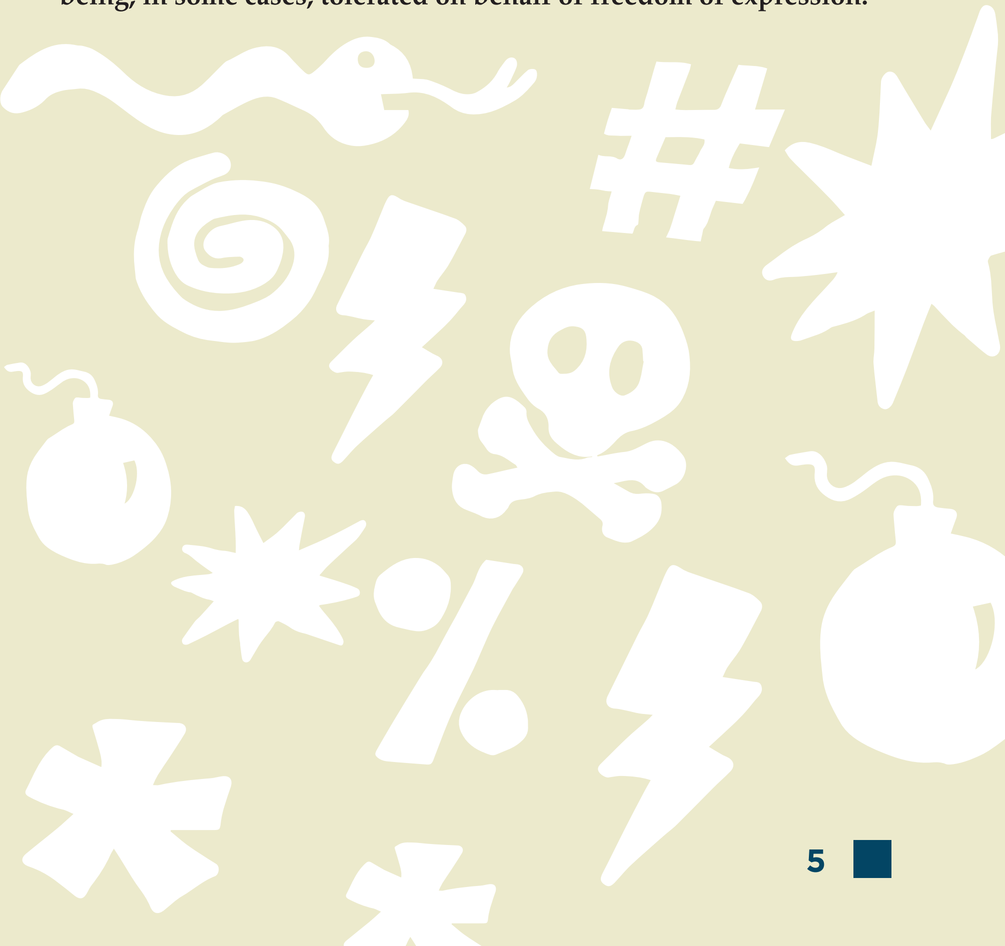
The individual who expresses the hate speech is herein referred to as the **"speaker"**; those to whom it is directed are **"the audience**", and those who are negatively evaluated by the hate speech are the "**target"**. A **vulnerable** group is so considered by being more likely to suffer violence or discrimination in comparison to other social groups.

Some examples of hate speech are attempts to dehumanize members of a historically discriminated group, comparing them to vermin or animals, or attempts to treat such a group as a threat to the well -being or wealth of the audience.

Hate speech is, therefore, an **"umbrella concept"** that encompasses several different forms of expression, grouped on account of similarities regarding their content, their target, the intention of their speakers and their potential effects. Those different forms of expression might be tolerated, sanctioned or regulated, depending on the evaluation of how severe they are.

Given its umbrella quality, several different legal provisions may be relevant to hate speech, which we call **sparse legislation.** In Brazil, for example, there are legal provisions that define as crimes certain forms of hate speech, even though this is not specified in the authoritative texts of statutes, nor in relevant provisions of international treaties that the country has adopted.

It should be emphasized that the concept of hate speech used herein is different from what is commonly found in Brazilian court rulings and legal scholarship. In these sources, hate speech is usually considered **illegal** in all its forms, a conclusion achieved by not separating the identification and evaluation stages of analysis. However, in order to understand the diverse aspects of the subject, as well as to organize discussions and comparisons to international sources, **we chose to de-naturalize this relation and to recognize the possibility of hate speeches being, in some cases, tolerated on behalf of freedom of expression.**



What makes **hate speech severe?**

Hate speeches have the potential to cause direct and indirect harm to the members of a vulnerable group. On the one hand, direct harm is psychological harm inflicted upon the members of vulnerable groups (who feel, for example, fear or anguish). On the other hand, indirect harm takes the form of acts of violence and discrimination that are a result of depreciation to the social standing of those vulnerable groups, which make them appear as undeserving of the same rights as other citizens. Depending on the theoretical perspective adopted, one or both different types of harm will be mentioned as justification for the regulation and sanctioning of hate speech.

In any case, hate speech is not to be confused with the mere offense to members of vulnerable groups, albeit the feeling of offense is generally present when it occurs. Even if a sanction against hate speech were justified based on the noted psychological harm, its severity must be greater than that of an offense.

It should be stressed that hate speech has the potential to **aggravate the vulnerability** of members of the target group, that is, to make acts of discrimination and violence against them more likely.

Even though it is difficult to accurately demonstrate a causal relation between hate speech and the occurrence of psychological harm or the increase of discrimination and violent acts, there are several studies that support this conclusion. It should be noted that it is often argued that the causal relation mentioned above does not stems from a single isolated manifestation, but from frequent manifestations which create a hostile environment that makes direct or indirect harm to the vulnerable group more likely. In this line of argument, one should determine how **severe** a given instance of hate speech is according to its lesser or greater contribution to a hostile environment.

What is the **Structured Set of Variables?**

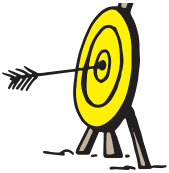
The **Structured Set of Variables (SSV)** is a set of variables that can be used to **identify** the occurrence of hate speech, **evaluate** its severity, and help deciding whether and how to **sanction** it. The SSV also considers that a **regulation** may be designed to create an environment that prevents hate speech or mitigate its effects.

Many of the variables were based on the potential that specific messages have of aggravating the target's vulnerability (indirect harm). Despite that fact, the SSV also contemplates the possibility of justifying a sanction or a regulation based on direct harms.

A theory about hate speech is paramount to provide an effective test for its identification, evaluation and regulation or sanctioning within a given legal order. Such a theory should select the appropriate variables according to legal tradition and historical context.

Identification

The initial step is to determine if a manifestation can be identified as hate speech, which depends on characteristics of the **(I) target**, **(II) message** and **(III) intentional context.**



I. Target

The target of hate speech needs to be a group considered vulnerable or an individual who is a member of such a group. The idea is that the condition of vulnerability is necessary to justify the special protection granted to these groups through the regulation or sanctioning of hate speech. It is therefore important to clearly delineate the concept.

A group is **vulnerable** if its members are more prone to suffering violence or discrimination, in comparison to other social groups.

Establishing vulnerability can be done in a variety of ways, all related to the empirical verification of a greater propensity to suffering violence or discrimination. Through data from sociological or historical research, it is possible to identify if the members of a group are victims of violence and discrimination more frequently than other individuals, or if they were victims of serious attacks in the past. Studies can describe the discrimination mechanisms that impact these groups (e.g. limitations preventing women from attaining specific positions in the corporate hierarchy). A legal analysis can also ascertain that specific groups are not holders of certain rights or find practical obstacles to the exercise of rights which other groups are entitled to (e.g. limitations on marriage between same sex individuals).

II. Message

The message conveyed by hate speech is the **negative evaluation** of the target.

This negative assessment can be direct or indirect. In the former case, the message explicitly says the target is less worthy of rights, opportunities, and resources. In the latter, the message negatively evaluates the target (i.e. they are criminals, vermin and parasites), in the hope that the audience will conclude by themselves that the group and its members deserve to be the target of discrimination or violence. Incitation against a specific group (i.e. "burn the indigenous people!") is also deemed to be an indirect negative evaluation, as it presupposes that the group is less worthy of rights, thus justifying the violence or discrimination.

Furthermore, the negative assessment of a vulnerable group can occur in non-discursive ways, such as using the Nazi swastika or the cross in flames from the Ku Klux Klan.

It should be highlighted that the existence of negative evaluations can be followed by discussions regarding how intense those evaluations are. Certainly, there are negative evaluations that are more virulent than others. Furthermore, a minimum degree of intensity might be deemed necessary to consider that a relevant negative assessment actually exists, one that allows for the identification of hate speech. Below this level, one would not be referring to hate speech, but merely to the expression of prejudice (someone could uphold, for example, that accusing refugees of speaking with unpleasant accents would simply be the manifestation of prejudice). This discussion on the degree of intensity of the negative evaluation is also relevant, subsequently, to assess the hate speech's severity, which also varies according to the message's content.

III. Intentional Context

In any instance of hate speech, the speaker negatively evaluates the target with the aim of establishing that she or they are less worthy of rights, opportunities and resources. This intention is perceived by the audience which attributes that meaning to the message.

However, some contexts may point to the existence of a diverse, different intention (i.e. an academic debate, humor) in such a way that the message will not be interpreted by the audience as having the goal of legitimizing discrimination or violence through negative evaluation. In that case, the message would lose the potential of aggravating vulnerability and would not cause direct harm, no longer being considered hate speech.

Evaluation

After identifying a manifestation as hate speech, it is necessary to determine its severity, with the aim of justifying a decision for sanctioning or tolerating it.

The evaluation's stage of analysis encompasses six categories of variables. These are organized based on the idea that, even after having identified hate speech, if its potential or risk of generating harm does not surpass a defined limit, there are no reasons for restricting it. In this case, freedom of expression would prevail. On the contrary, once the limit has been surpassed, the evaluation is relevant to determine the form of sanction, whose severity will increase accordingly with the seriousness of the hate speech.

One of those variables (IV. Situational Context) is used to establish the minimum degree of severity that needs to be surpassed for sanctions to be considered necessary, while the other five are used, each in its own way, to evaluate its actual severity, that is, the potential of a given hate speech for inflicting harm or contributing to a hostile environment.

There are two notions that permeate all the other categories (V to IX) and help to understand the meaning of the variables: the speech's **reach**, that is, its capacity to impact a large number of people, and its **persuasive impact**, that is, it's ability to generate sufficient impact on each individual it reaches, to the point of changing their mindset and behavior.

IV. Situational Context: The tolerance level to the hate speech can be greater or lesser in specific situations, according to some justifying reasons. In other words, some relevant situations allow for certain types of hate speeches to be tolerated, due to other values that need to be protected, depending on their severity.

> **Ex**. If a hate speech is delivered at religious preaching, given the special protection granted to the freedom of religious expression, a great degree of severity would be demanded to justify imposing a sanction. The same applies to hate speech conveyed within the context of a political debate. Such speech could be considered tolerable, due to the special importance attributed to freedom in political expression.

V. Speaker: Who delivered the hate speech? How do some characteristics of the speaker influence its potential to harm?

> **Ex**. The hate message of a religious leader or that of a movie star can draw the attention of a great number of people (reach) and have significantly more persuasive impact among their followers. The same is true for a speaker that holds some sort of political or economic power over its audience.

VI. Audience: Who was the hate speech meant to be heard by? What characteristics of the audience can make them susceptible to the persuasive impact of the message?

> **Ex**. When the audience already fear or holds a grudge against the target group and/or holds the necessary tools to act in a violent way against the target (the case of armed and organized groups), it is more likely that the hate speech geared towards this audience will escalate into acts of discrimination and violence.

VII. Message's Vehicle: What are the means through which is the message disseminated? Which characteristics of this mean could grant wider reach and more persuasive impact to the message?

Ex. A hate speech disseminated through a television program broadcasted on prime time or through a popular TV channel has more potential to do harm than one disseminated through pamphlets or posters by an individual on streets that are not busy.

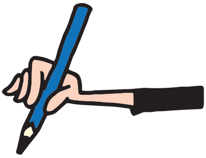
VIII. Socio-Historical Context: In which social and historical context is the hate speech being delivered? How does that context increase or decrease the risk of an outbreak of violent and discriminatory actions?

Ex. Hate speech can catalyze violent actions with greater ease when the groups involved (target and audience) historically compete, either for resources, or due to religious discrepancies or political divergence. The competition creates feelings of resentment, rivalry or even of vengeance that may draw these people closer to turning discourse into action.

IX. Consequences: Which concrete and verifiable consequences of the hate speech can be observed? What do they inform, retrospectively, about the severity of the speech?

Ex. In some cases it may be possible to demonstrate, with a high degree of certainty, that discriminatory or violent behaviors were committed due to a particular instance of hate speech. This happens, for example, when these acts take place soon after the speech is delivered, or when its authors claim having acted because of a specific hate message conveyed.

Sanctions and Regulation



The identification and evaluation of hate speech, despite involving a series of inherent difficulties, are preliminary to the **decision-making stage** that consists in choosing between freedom of expression or sanctioning of a specific manifestation identified as hate speech. Herein we list several of the possibilities for reacting to a serious instance of hate speech.

A wider analysis of the hostile environment created by hate speech manifestations of a given society may also be considered, resulting in alternative forms of regulation. It is not only about sanctions because the ways of dealing with the problem do not necessarily involve punishing the speaker, nor even suppressing the speech. These alternatives are **policies for the prevention of hate speeches and their effects**, as well as the **emission of counter-speech,** with the goal of creating a social environment that will reject hate speech. Prevention policies have a special characteristic since they do not refer to a particular manifestation.

X. Prevention Policies: Measures that can prevent the occurrence of a hate speech or mitigate its effects, mainly through the limitation of its reach or its persuasive impact.

Ex. Some measures to limit the reach are already adopted by social media platforms such as Twitter, for example. There are measures to restrict the visibility of a message, its circulation, its appearance in search results, its appearance in users' timelines etc. Moreover, the persuasive impact can be limited by measures that create empathy between the potential audience and the possible targets groups of hate speech, which can be done, for example, through artistic manifestations in different media.

XI. Counter-speech: It is about delivering a speech that is contrary to a specific occurrence of hate speech, with the intention of contesting such speech, by presenting arguments or different versions of the facts.

Ex. The counter-speech can arise spontaneously, as a response to a newspaper article or as a retort to a post in a social network. It can also be engendered through a legal route, as in cases in which, based on the occurrence of hate speech, it is requested that an educational program be produced to elucidate and reaffirm the rights of the target group.

XII. Removal: These are measures that involve removing hate speech from circulation after it has been released.

Ex. This removal can be judicially ordered, but may also be done through the route of self-regulation or private regulation, such as through moderation in a public forum on the Internet, or through the practice of social media platforms limiting content based on the enforcement of their Terms for Use.

XIII. Censorship: It forbids the utterance of a certain instance of hate speech. Here, we consider the hypothesis of censorship exercised by private entities, such as social network platforms, considering that previous censorship by public powers is expressly prohibited by the Brazilian Federal Constitution.

Ex. There is the possibility of implementing filters in social networks, which detect and prevent individuals from posting extremist groups flags or symbols.

XIV. Civil Penalties: Hate speech may be considered a tort, and the speaker or even third parties may be held liable.

Ex. There are many legal cases where it has been alleged that hate speeches caused collective damages, and indemnification for such damages were claimed judicially. Social media platforms may also be obliged to pay for damages caused by content aired by their users, whenever they do not comply with a judicial order to remove it.

XV. Criminal Sanctions: Some forms of hate speech can be deemed serious enough to justify a criminal sanction, and may be considered crimes by Brazilian Law.

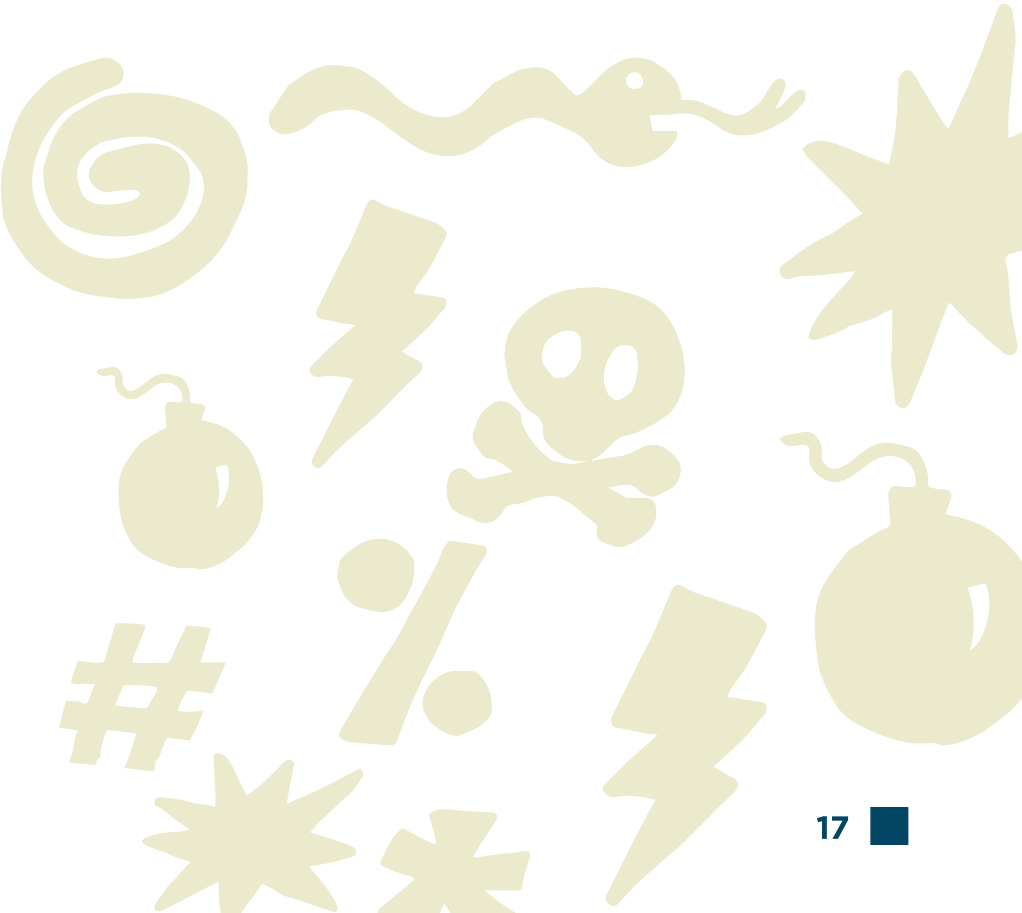
Ex. The Brazilian law criminalizes behaviors that may be deemed as hate speech, despite not containing the term “hate” in its text. It is the case of the crime of practice, and incitement to discrimination or prejudice due to race, color, ethnicity, religious or national origin (article 20 of Law 7.716/89) and that of insulting based on elements that refer to race, color, ethnicity, religion, origin or condition of an elderly person or one bearing a deficiency (article 140, § 3º of the Penal Code).

XVI. Administrative sanctions: Hate speech can be sanctioned administratively by public institutions.

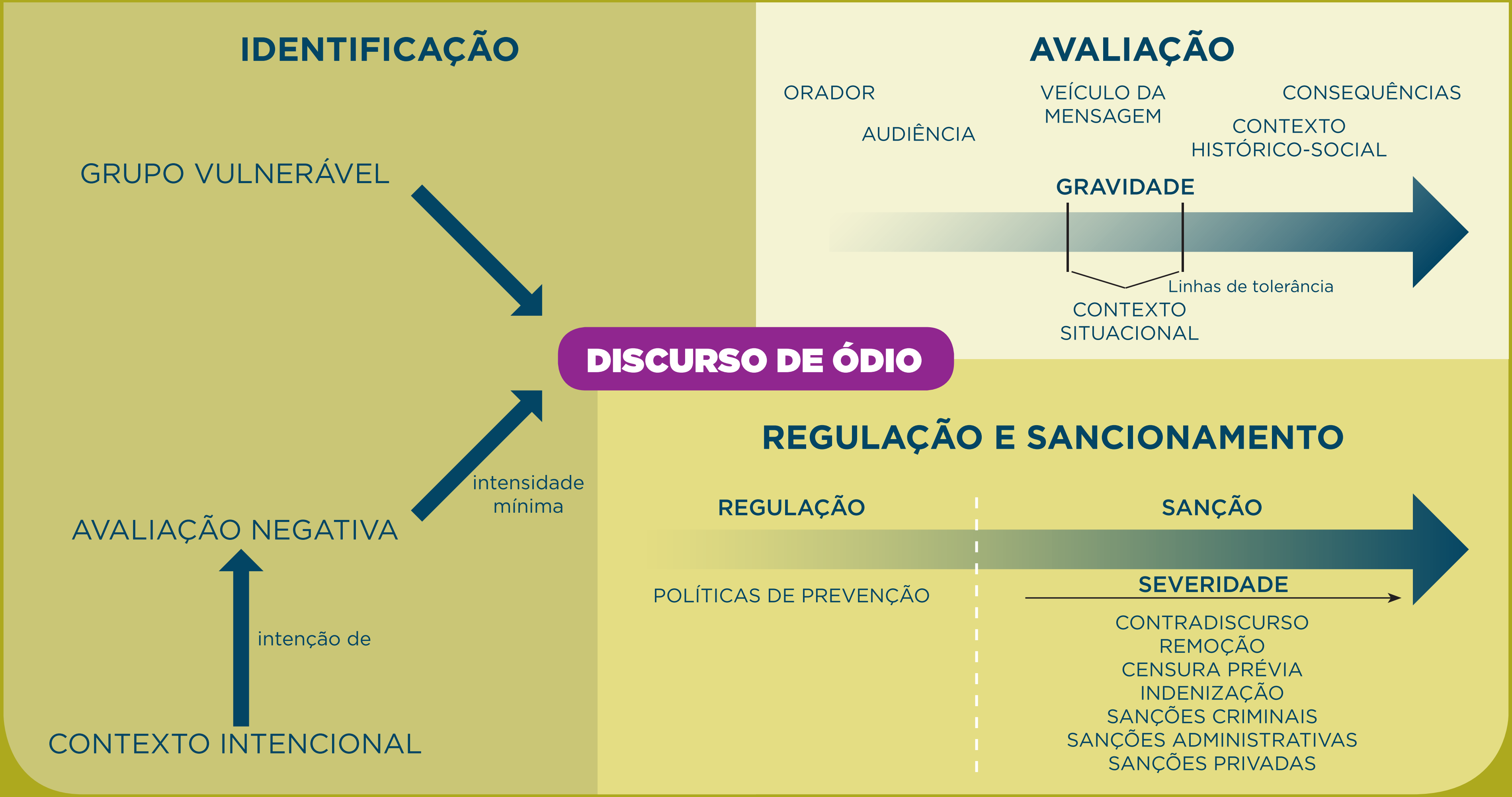
Ex. An open television channel, as a public service provider, is subject to the administrative regulations, and may have the concession cancelled or not renewed because of acts deemed to be abusive (i.e. fostering discriminatory campaigns, in the case of article 53 of the Brazilian Telecommunications Code).

XVII. Private Sanctions: A series of sanctions may be applied by private institutions, as a result of their self-regulatory powers. The most significant examples of those institutions would be social media platforms, educational institutions, and corporations.

> **Ex.** Corporations have the choice of setting forth their own codes of conduct, where they may provide for disciplinary sanctions to be enforced on undesirable behaviors in the work environment. For example, The Code of Conduct of a company may state that it will not tolerate harassment, insults, threats, or other undesirable behavior, based on characteristics that define vulnerable groups.



Infográfico da **Matriz de Variáveis**



IDENTIFICAÇÃO

GRUPO VULNERÁVEL

AVALIAÇÃO NEGATIVA

intensidade mínima

intenção de

CONTEXTO INTENCIONAL

DISCURSO DE ÓDIO

AVALIAÇÃO

ORADOR

AUDIÊNCIA

VEÍCULO DA MENSAGEM

CONSEQUÊNCIAS

CONTEXTO HISTÓRICO-SOCIAL

GRAVIDADE

Linhas de tolerância

CONTEXTO SITUACIONAL

REGULAÇÃO E SANCIONAMENTO

REGULAÇÃO

POLÍTICAS DE PREVENÇÃO

SANÇÃO

SEVERIDADE

CONTRADISCURSO
REMOÇÃO
CENSURA PRÉVIA
INDENIZAÇÃO
SANÇÕES CRIMINAIS
SANÇÕES ADMINISTRATIVAS
SANÇÕES PRIVADAS

Work Group

Fernando Lottenberg
President of Conib

Rony Vainzof
Director Secretary of Conib

Sergio Napchan
Director General of Conib

Alexandre Pacheco da Silva
Marina Feferbaum
Coordinators of the CEPI/FGV

Victor Nóbrega Luccas
Researcher Coordinator of the CEPI/FGV

Fabricio Vasconcelos Gomes
João Pedro Favaretto Salvador
Researchers of the CEPI/FGV

graphic project: Dagui Design – illustrations: Ivo Minkovicus

FGV DIREITO SP
CENTRO DE ENSINO
E PESQUISA EM INOVAÇÃO

CONIB
Confederação Israelita do Brasil