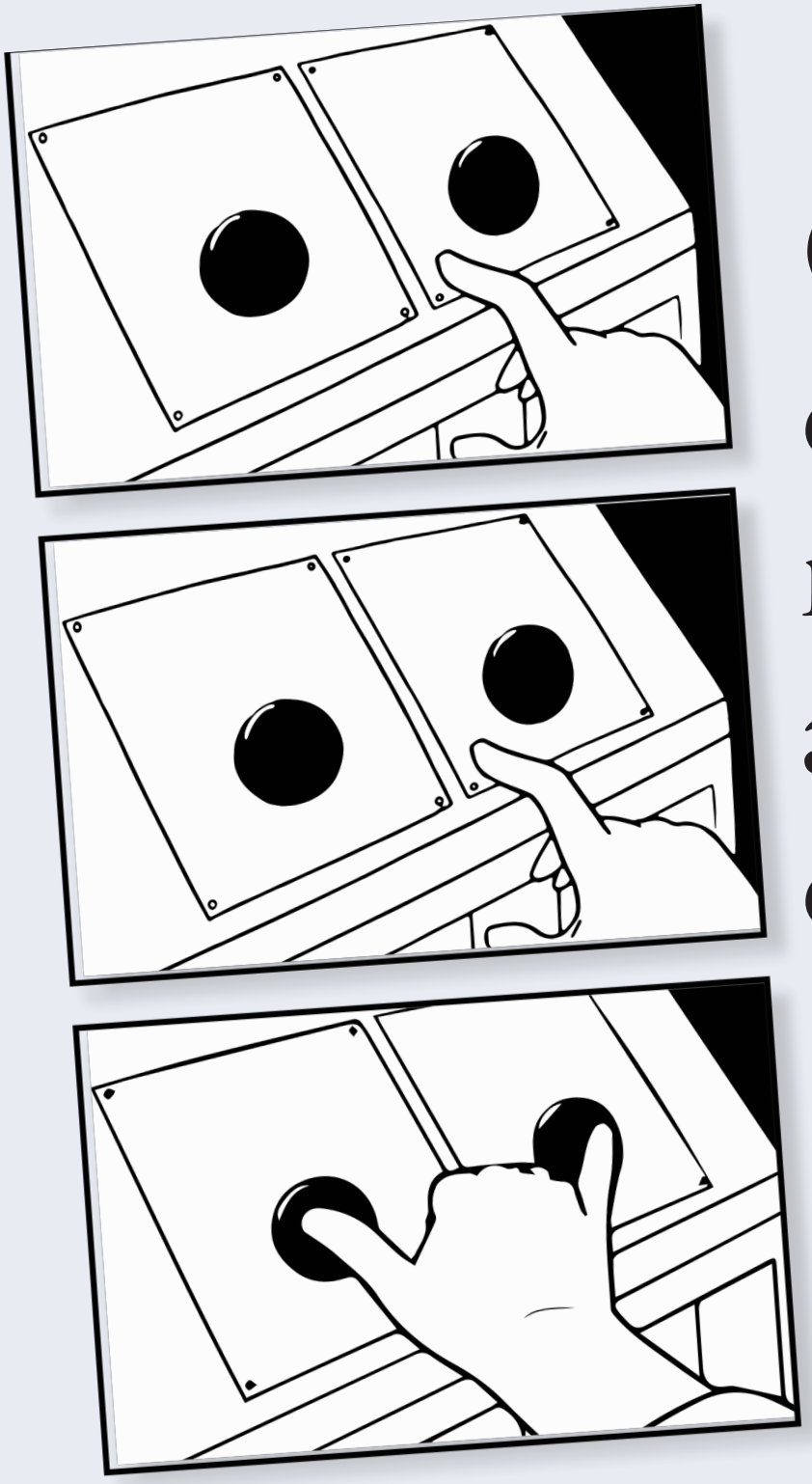
ISSN 2993-1908

SPECIAL COLLECTION OF THE CASE LAW ON FREEDOM OF EXPRESSION



Case law on content moderation and freedom of expression

Global Freedom of Expression
COLUMBIA UNIVERSITY

Global Freedom of Expression
COLUMBIA UNIVERSITY

Case law on content moderation and freedom of expression

JUNE 2023

Global Freedom of Expression
COLUMBIA UNIVERSITY

Credits

Collection directors

Lee C. Bollinger
Catalina Botero-Marino

Author

Erik Tuchtfeld [*], *Research Fellow at the Max Planck Institute for Comparative Public Law and International Law, Heidelberg, Germany*

Design

Laura Catalina Ortiz, *illustrator*
Lourdes de Obaldía, *layout and graphic designer*

Special thanks and acknowledgements

The Directors, Editors, and Authors of the present collection would like to recognize and express their gratitude to all the people whose efforts and talents made the collection a reality. These publications were only possible thanks to the analysis and selection of cases for the database by a wide number of experts and contributors collaborating with Columbia Global Freedom of Expression. The case briefs presented in this collection reproduce the analysis of the cases published in our database, which was only possible due to their invaluable contribution.

Special thanks to Juan Manuel Ospina for reviewing and providing valuable comments on the paper.

[*] The views and opinions expressed in this paper reflect only the position of its author and do not necessarily reflect the position of Columbia Global Freedom of Expression.

Table of Contents

Table of Contents

I. Introduction

The state of freedom of expression online is substantially shaped by the moderation of content on dominant social media platforms. Their private community guidelines are the legal framework for discussions in the digital realm. Effectively, they constitute global law applying to billions of individuals, set by a few companies of the Global North. This raises fundamental questions on how to deal with private power. The answer varies substantially, depending on the legal culture: Some, in particular the United States, emphasize the principle of private autonomy. When users choose a particular platform, they accept its terms and conditions. If they are unbearable for the users, they can choose another platform. Thus, it is not on the state to impose any rules on this contract of private, autonomous parties. Others understand the issue of private power as an antitrust problem. Too few companies and quasi-monopolistic structures hinder the free market. The lack of competition, reinforced by network effects, leads to walled gardens, where consumers are left with no other choice than to pick one of the few existing networks and accept the rules of a "benevolent dictator". In consequence, only breaking up global social media companies or, at least, setting up interoperability obligations for them could improve the situation. A third school of thought addresses the accumulation of private power as a human rights problem. While nation-states are the main and "classical" addressees of human rights obligations, powerful private companies are bound as well, although to a lesser degree. This approach is the one analyzed in this Special Collection paper. Claimants all around the world have initiated legal actions based on the assumption that a particular act of content moderation—such as the (non-)removal of content, account's suspensions or deletions, among others—, violated their human rights.

However, the application of human rights to content moderation does not necessarily facilitate its exercise in practice. "Chilling effects" as a result of too restrictive laws are well-known to the human rights discourse. Individuals alter the way they exercise their rights when they fear sanctions. It is not even necessary that a concrete action is in fact punishable, it is enough for individuals to be uncertain of the consequences they would face, to stop them from doing it. This could be understood as a call to moderate content as little as possible to enhance freedom of expression. But it's not that easy. For example, hate speech often attacks a person's dignity, which is also protected by human rights law. Beyond that, not only "chilling effects" but also "silencing effects" can be observed. Humans, in particular women and members of vulnerable groups, are leaving the online discourse because of the hate and aggression they experience there. Participation in the digital public space—mainly made up of social media platforms—lies at the core of what freedom of expression should enable people to do. Thus, the balancing of different rights is a difficult challenge for courts that review content moderation decisions taken by social media platforms in the first place.

In the process of elaborating this paper, more than a hundred cases on content moderation from all around the world have been considered. Not all of them have made it into this paper. Due to the broad scope of the topic of content moderation, some limits had to be drawn: This paper does not include cases in which *individuals* fight over the legality of a particular statement. While such cases often take place on social media platforms, they do not concern content moderation in a narrow sense. They do not delve into *platforms'* obligations to remove (or reinstate) content, rather –just like in an analog environment– they concern general issues of defamation law and freedom of expression. Also, copyright cases are not included. They do address the obligations of social media platforms to keep their platforms free from illegal content. However, their legal regime is highly specialized and outside the scope of inquiry of this paper.

The "classical cases" of content moderation dealt with in this Special Collection concern the platforms' handling of hate speech, political issues, defamation, and disinformation which are, due to a lack of a special legal regime, in most cases solved by applying general principles of law.

Not only social media platforms (in a narrow sense) are exercising such content moderation. While

they are characterized by their users' ability to upload content at their own discretion, search engines, on the other hand, do not allow such uploads. However, the very core of what a search engine does is to allow the exploration and curation ("moderation") of content it finds on its own on the internet. As search engines are pivotal for finding (possibly embarrassing or even defamatory) content on the internet, claims directed against these companies have been brought forward to stop the distribution of information that was considered wrong or hateful. Since a comprehensive overview of search engines' obligation to remove content that is incorrect, outdated, or simply not sufficiently relevant, while balancing the rights of the individual in question and the public, already exists in the *Special Collection of the Case Law on Freedom of Expression: Does our past have a right to be forgotten by the Internet? Case Law on the So-Called Right to Be Forgotten*, cases concerning search engines' obligations in this regard have not been added to this Collection.

Finally, this Special Collection is naturally limited to the cases which are already included in the Global Freedom of Expression Case Law Database. The database hosts more than 2000 cases in total and, with that, nearly all the leading cases on freedom of expression in recent years. While the database relies on a global network of contributors, cases from the United States, Latin America, and Europe are most prominent. Thus, this Special Collection does not claim to give an all-encompassing overview of all the relevant jurisprudence on content moderation worldwide. However, the carefully chosen sample aims to cover and contextualize the most important legal debates that courts are currently facing on this matter. It also tries to showcase the different pathways taken by judicial bodies when solving these issues.

This Special Collection only concerns cases issued by state courts dealing with matters of content moderation. Of course, it is social media platforms which, as part of their daily business, are issuing by far the most decisions on content moderation. They do so a thousand times per day. Some of them, like Google and Meta, have decided to commit themselves to the United Nations Guiding Principles on Business and Human Rights (UNGPs), and self-regulatory bodies, such as Meta's Oversight Board. This Collection does not include their decisions. For more information on them, see the Special Collections of the Case Law on Freedom of Expression: Meta's Oversight Board Cases, and The Decisions of the Oversight Board from the Perspective of International Human Rights Law.

The jurisprudence presented in this paper is divided into three main sections: Cases against intermediaries, cases against public officials, and cases dealing with the actions taken by states to enforce a particular kind of content moderation on private social media platforms.

The first section includes cases lodged by individuals against intermediaries, mainly social media platforms. Claims by individuals compelling platforms to "carry" specific content are often based on an alleged violation of freedom of expression materialized by the removal of content or the suspension of accounts. Thus, their success is strongly linked to the position of the respective legal system towards the (indirect) effects of fundamental rights on private relationships. When it comes to the removal of content, the responsibility of the platforms for the content of third parties —the users— often comes into question.

The second section concerns cases against government officials administering social media pages. Here, the distinction between private and official pages is especially challenging. When do private social media profiles "transform" into public ones? Can public fora exist within the private "property" of a social media platform?

Lastly, the third section examines cases dealing with governmental responses to a supposed insufficient moderation of content by social media. Measures taken in these instances range from voluntary agreements to hard bans of specific social media platforms.

II. Cases against intermediaries

Court decisions on content moderation mainly deal with the obligations of private enterprises. How free are they in their decisions of (not) removing content? What kind of boundaries – if any at all – do criminal law, contract law, and even fundamental rights set? The analysis of court decisions from around the world show that the answer to these questions very much depends on the particular legal system.

The selection of cases in this section is divided into three main categories: (1) Cases in which users asked for the reinstatement of posts or accounts. In these cases, the decisive element of the judicial system comes to its stance on the indirect effect of fundamental rights on private parties. Most legal systems accept that fundamental rights can also be taken into consideration in contractual relationships between private parties, at least in cases where there is a significant imbalance of power. Thus, the users' right to freedom of expression *vis-à-vis* the platform's right to decide which content is allowed on its space is balanced. USA courts, however, have rejected such approaches consistently and emphasize their "state-doctrine" in which the state is the sole addressee of fundamental rights obligations. (2) Claims against the platforms to remove content that is deemed unlawful. This concerns one of the most important legal issues regarding the structure of today's internet: the liability of hosting providers for illegal content shared via their services. Some of these cases not only request the removal of content (one could call this the "primary claim") but also damages for the failure of doing it in time ("secondary claim"). (3) The cases in the third category are insofar different, as they do not concern specific content (or accounts) but ask for information on how the moderation of content is organized. They are investigative in nature and address mostly systemic issues rather than concrete ones.

1. Claims to reinstate content or accounts

Obligations to reinstate content or accounts have been called "must-carry" orders. In the USA context, these refer to a long-established doctrine that obliges private entities to "carry" material by other private entities, often due to its monopolistic position and/or its public function. While such claims remain to this day unsuccessful in the United States, when it comes to content or accounts on social media platforms, courts in other nation states have granted them. In particular, German courts have underscored the importance of social media platforms for the public discourse, its quasi-monopolistic position, and the consequences that platforms' decisions have for the exercise of fundamental rights. As many states have an established doctrine of (indirect) effects of fundamental rights on private entities, courts have had little problem applying fundamental rights when assessing the concrete contractual obligations of a social media platform. What they struggle with is not the binary question *of whether* fundamental rights are applicable, rather to what extent a private entity is bound to fundamental rights, and how conflicting rights shall be balanced.

On the other hand, in the United States, the fundamental rights enshrined in the Constitution only establish obligations towards the state. If an action cannot be attributed to the state, fundamental rights are, in principle, of no concern (state-action doctrine). Thus, the arguments and decisions of courts in the United States differ vastly from other jurisdictions when it comes to content moderation.

a. Content or accounts of individuals

The jurisprudence from Germany can be categorized in two main schools of thought: Courts that have assigned state-like obligations to social media platforms to respect freedom of expression, and courts that emphasize social media platforms' own rights (e.g. the right to conduct a business) and focus on non-arbitrariness by introducing procedural obligations. This conflict has been decided (at least for the moment) by the Federal Court in favor of the latter.

One example of the first school of thought, in which it was held that Facebook had to respect its user's freedom of expression, is the decision in favor of the right-wing politician Heike Themel by the Higher Regional Court Munich, **Germany**, from August 2018. Themel's account was suspended after commenting in response to a slur—in an online debate below a Facebook post—that she couldn't argue with another user as they were "[argumentative] unarmed]" and it "wouldn't be fair" from her side. The Court examined Facebook's community guidelines and held that Facebook's unilateral prerogative to decide which posts or comments breached its community guidelines was in violation of German contract law. Instead, the interests of both parties that agreed to the contract had to be taken into account. The Court emphasized that the purpose of Facebook, as a social media platform, was to give its users a "public marketplace" for exchanging views and opinions, and noted that in the context of the right to freedom of expression, permissible expressions of opinions cannot be deleted from the platform. As Themel's comment did not constitute criminal hate speech according to the Court, Facebook was ordered to reinstate her comment and grant her access to her Facebook account.

Similarly, the Regional Court of Cologne, **Germany** (2018), found that Facebook could not block a user's account for harsh comments, or remove comments related to specific incidents, which do not constitute incitement as penalized by the German Criminal Code. The user had criticized asylum seekers in Germany using degrading terms. When interpreting Facebook's Community Standards section on hate speech, considering the right to freedom of expression, the Court underscored the open and vague wording of the language used by the company. Thus, for the Court, an average user would be right to assume that Facebook offers everyone the opportunity to disseminate facts and opinions on all conceivable topics, including comments with drastically formulated value judgments, as long as they are non-punishable under German law.

In January 2020, the Higher Regional Court of Munich, **Germany**, reiterated once again its position that Facebook must not sanction any comment which is covered by an individual's right to freedom of expression as guaranteed in the German Constitution. It emphasized that this right is not absolute but limited by criminal law and by conflicting fundamental rights, such as the right of personality of other individuals. Any platform's decision qualifying a post as forbidden hate speech is fully reviewable by courts. The concrete statements in question, which referred to Muslim refugees as "invaders", were—in the opinion of the Court—still protected by freedom of expression and, thus, couldn't serve as the basis for suspending the user's Facebook account. However, since Facebook's state-like obligation to respect freedom of expression derives from its purpose as a forum for the general exchange of information and opinions, companies could avoid such duty by opening forums that are, from the beginning, tailored for a particular purpose.

A different approach was taken by the Regional Court of Heidelberg, **Germany** (2018), which held that Facebook's Community Guidelines met the requirements of transparency and non-discrimination under German contractual law, and adequately considered users' rights to freedom of expression. The Court focused on the nature of fundamental rights as imposing negative obligations addressed towards the state, not other private entities. While they deploy (indirect) effects on the contractual relationships between private parties, this does not mean that Facebook must respect aggressive speech to the same extent as the state does. Instead, Facebook must consider its user's right to freedom of expression, but it also has a wider margin of appreciation than the state when limiting it.

This stream of argumentation was shared by the Higher Regional Court Stuttgart, **Germany** (2018), when assessing the suspension of a user account following anti-migration comments. The Court held that the suspension was lawful and confirmed that the platform's community guidelines met the requirements of transparency and non-discrimination under German contractual law. While Facebook was bound to respect freedom of expression to a significant degree, especially "given the respondent's dominant position in the market for social networks and the great importance of freedom of expression in a democratic constitutional state", Facebook itself can invoke fundamental rights such as "virtual domiciliary rights" and entrepreneurial freedom. Also, as Facebook is liable to incur penalties for not removing criminal comments under the German *NetzDG*, it is permitted to avoid such responsibility by removing questionable content.

Finally, the **German** Federal Court decided on this controversy on July 2021. The Court rejected the view by which Facebook is bound by fundamental rights to the same degree as the state, since its market-dominating position "cannot be equated with the monopoly position of state-owned companies in the area of public services". Instead, the right to freedom of expression of users must be balanced with Facebook's own rights. Thus, Facebook is "in principle entitled to require the users of its network to comply with certain standards of communication that go beyond the requirements of criminal law". However, there must be an objective non-arbitrary reason for any sanction, and procedural safeguards necessary for the clarification of a case must be put in place. In particular, Facebook is obliged to inform its users immediately about the reasons why it removed a post, and has to grant them an opportunity to respond, followed by a new decision. In the case of the (planned) suspension of a user account, the user even needs to be heard before the suspension takes place.

In essence, the Court did not express clear limits on how far Facebook could deviate from the obligations that freedom of expression imposes on the state. Nonetheless, it granted "fair trial-rights" to Facebook's users, such as the right to be heard and the principle of non-arbitrariness, in order to achieve a certain balance between conflicting fundamental rights. Since Facebook's terms of service did not include such safeguards, they were declared null and void.

In opposition to the German jurisprudence, that focused on the degree in which fundamental rights' obligations apply to social media platforms, cases before US courts dealt with the question of whether any liability at all can be assumed. All judgments deny this in a coherent manner. Fundamental rights, in the US legal system, are addressed towards the federal government and the states, and they do not alter the relationship between private parties. Moreover, Section 230 of Title 47 of the United States Code, protects any platform from liability, as long as it satisfies a three-step-test: (1) The platform must be a provider or user of an interactive computer service; (2) the relevant information or content must be provided by another information content provider; and (3) the complaint must seek to hold the platform liable for its exercise of a publisher's traditional editorial functions —such as deciding whether to publish, withdraw, postpone or alter the content.

For example, in *Mezey* (2018) the **US** District Court for the Southern District of Florida dismissed a lawsuit filed against Twitter for suspending the claimant's account without providing any reasonable justification. The Court held that Twitter was protected under Section 230, as it transmits, receives, displays, organizes, and hosts content and, thus is an "interactive computer service". Also, the information it hosted was provided by another information content provider (its users), and the platform's activities could be boiled down to deciding whether to exclude material that third parties seek to post online, which falls under the definition of a "traditional editorial function".

This understanding of Section 230 was also confirmed by the **US** Court of Appeal for the State of California in *Twitter v. The Superior Court for the City and County of San Francisco* (2018), after the platform suspended two accounts for violating the "Violent Extremists Group Rule". The Court referred to case law which emphasized that Section 230's immunity is "broad and robust". In consequence, Twitter had to be understood as a publisher and was barred from any liability by the aforementioned provision.

In *Cox* (2019), the **US** District Court for the District of South Carolina Division emphasized once more Twitter's qualification as a publisher under Section 230 and, thus, dismissed an action filed by a Twitter user whose account was suspended after publishing a tweet criticizing Islam. The Court also examined an alleged breach of contract by Twitter for requiring the user to delete content to regain access to his account. The contract both parties entered, Twitter's Terms of Service Agreement, however, reserved Twitter the right to remove content that infringed its agreement and to suspend or cease to provide the user with all or part of its services. To accept Twitter's unilateral right to decide on content's compliance with its own Terms of Service contrasts sharply with German jurisprudence, where even courts which were reluctant to apply fundamental rights' obligations to social media platforms stressed that it is not on the platform to decide de-

finitively which content violates its Terms of Service, but that such a decision is fully reviewable by a court.

In *Williby*, the **US** District Court for the Northern District of California dismissed a First Amendment claim against Facebook in June 2019. The claimant's contention —that Facebook was a public forum for speech and was consequently bound by the First Amendment's speech guarantees— was disregarded by the Court as the First Amendment only applies to governmental abridgments of speech and not to alleged abridgments by private companies like Facebook. The Court held that Facebook did not engage in an activity that the government has traditionally and exclusively performed and, thus, did not qualify as a state actor. Instead, as a private entity, Facebook was entitled to exercise editorial discretion over the speech and speakers in its forum.

More recently, in the context of the global COVID-19 pandemic, the issue of disinformation, in particular relating to health information, gained traction. Social media platforms committed themselves to fight disinformation and even enacted specific policies on COVID-19 and health related matters. Also, Facebook, Twitter, and Google signed, among others, the EU Code of Practice on Disinformation —developed under the patronage of the European Commission, which serves as a voluntary framework of industry self-regulation.

The question on the platforms' discretion to moderate health information also arrived at the courts. In *Van Haga*, the District Court of Amsterdam, in the **Netherlands** (2021), dismissed a claim by a Dutch Member of Parliament after YouTube removed a video containing an interview with him about the national Covid-19 measures. The claimant argued that the removal of the video was in breach of the contract between him as a user and YouTube. While the Court stressed that private law norms must be interpreted in light of fundamental rights, the mere fact that YouTube has the ability to reach huge audiences, or could even be said to have a "near-monopoly position", is insufficient to force it to tolerate each and every expression made by its users. Since YouTube's policy was based on scientific consensus —as communicated by the World Health Organization and national health authorities, as well as on the European Commission's call for help to combat misinformation about Covid-19 (June 2020), and on the Code of Practice on Disinformation (2018)—, the company, the Court opined, had responded to governmental instructions, and could therefore not be said to have acted unreasonably. Also, the claimant had, as a Member of Parliament, "sufficient possibilities to express his views, especially on the 'platform' explicitly intended for that purpose: the House of Representatives."

b. Content or accounts of political organizations

Of a particularly sensitive nature is the suspension of political organizations from social media platforms. They are often associated with political candidates who compete for the support of citizens in elections and take active part in the public debate. To address voters, they rely on modern communication tools, such as social media platforms. When compared to sanctions against individuals, measures against organizations often provoke a particularly fierce public reaction.

One of the first landmark decisions on such matters, occurred when **Germany's** Constitutional Court issued a preliminary injunction ordering Facebook to restore access to the suspended user account of the right-extremist party "Der III. Weg"—just shortly before the 2019 elections to the European Parliament. The party's Facebook account was suspended for sharing a link to an article in which asylum seekers were associated with violence and criminal offences. Facebook considered the article was hate speech and violated its community standards. The platform disabled the party's account, so that the account and its content were no longer available. The Constitutional Court held that it was essential for the party to have access to its Facebook page to disseminate its political opinions and discuss them with its users, until the elections were carried out. Facebook had a key position within the social networks in Germany and therefore played an important role in election campaigns. The Court highlighted the existence of many complex legal questions in this case which it did not aim to resolve definitely at this stage (preliminary injunction). Instead, it

applied the so-called "Doppelhypothese" test (dual hypothesis), developed for preliminary injunctions in constitutional proceedings, which requires the Court to balance the possible consequences when a preliminary injunction is granted, but the final claim is ultimately rejected in the principal proceedings, and vice versa. The Court concluded that the consequences that would arise if the applicant was denied access to its Facebook page clearly outweighed the consequences that would arise if the respondent was temporarily obliged to restore access, even if in the main proceedings it was ultimately decided that restricting access was lawful.

One year later, the Higher Regional Court of Dresden, **Germany** (2020), had to decide whether Facebook was permitted to block access to its platform to users who associated with "hate organizations", as determined by Facebook —in this case the right-wing association "Ein Prozent". The Higher Regional Court emphasized that social networks are in principle free to exclude "hate organizations" and their supporters under their terms of use, as long as these exclusions are not arbitrary, and take into account the users' fundamental rights and the economic effects of a permanent exclusion. However, it pointed out that the classification of a "hate organization" as such is fully reviewable by the Court, which must consider all individual circumstances. In the case at hand, the Court considered that the objective conditions laid out in Facebook's terms and conditions for considering "Ein Prozent" as a "hate organization" were met.

Similar circumstances led to a remarkable divergence in two decisions issued by the Court of Rome, **Italy**. In February 2020, the Court held that Facebook was entitled to suspend the accounts of the neo-fascist party "Forza Nuova" and its members. After their accounts were suspended due to racist, fascist and xenophobic comments, Facebook users approached the Court, arguing that the company's actions violated their right to freedom of expression. The Court conducted an extensive overview of national and international legislation and case law. It observed that international law, when assessing the limits to freedom of expression, does not allow for any protection of hate speech or discrimination. Furthermore, it analyzed jurisprudence by the ECtHR and as well as recent initiatives by EU institutions to combat hate speech in the digital sphere —such as the Code of conduct on countering illegal hate speech online. The Court concluded that the suspensions were in accordance with Facebook's Terms of Service and the Terms themselves were lawful, as hate speech is not protected by the right to freedom of expression, and Facebook is both permitted and required —under the EU's Code of Conduct— to take action against hate speech on its platform.

However, roughly two months later, the Court dismissed Facebook's appeal against a preliminary injunction ordering the social network to reactivate the account and restore the pages of the Italian neo-fascist party "CasaPound". Facebook claimed that violence and racism as enacted by the party's members amounted to an infringement of its Community Standards, thus allowing the company to disable the party's account. The Court went substantially deeper into the merits than the German Constitutional Court did in the *"Der III. Weg"* case, and stressed that the limits of freedom of expression are set by criminal law and law of associations, enacted by the Italian legislator. It held that Italian law does not prohibit neo-fascist associations in themselves, unless they attempt to reconstruct the Fascist Party of WWII. Facebook's Terms of Use, then, must be interpreted and applied in compliance with the principles of freedom of thought and association as guaranteed by the Italian constitution. Thus, Facebook had unduly terminated the contractual relationship on the basis of a constitutionally protected act of freedom of thought.

While the decisions by German and Italian courts differ in their individual results, they follow the same line of argument: A social media platform has to consider the fundamental rights of its users when sanctioning them. The terms and conditions are not at the complete disposition of the platform, but constitute a legally binding instrument which can be interpreted by a court to assess its "true" meaning. Following the doctrine of the horizontal effects of fundamental rights, the courts need to take into account the freedom of expression of the users when confronted with a quasi-monopolistic private entity like Facebook or Twitter. This view is not shared in the **United States**. Courts there have reiterated again and again that social media platforms are not bound by fundamental rights, rather they are protected from all claims by Section 230.

Claimants in different cases have invoked in particular (but unsuccessfully) the notion that social media platforms are by now "public forums" as conceptualized by the US doctrine.

For example, in *FAN* (2019), a **US** District Court in California ruled that Facebook's conduct in removing a Russian news distribution organization's Facebook account and page did not violate the First Amendment's protection of freedom of expression, and that its conduct was immunized from civil suit by US legislation. Facebook had shut down the page following the 2016 United States presidential election on the grounds that it was one of the "inauthentic" accounts that had allegedly sought to inflame social and political tensions in the United States. The Court held that Facebook had not violated the First Amendment as it is neither a public forum nor do its actions amount to state action. Further, the Court held that since Facebook is a provider of interactive computer services it had immunity under Section 230. Regarding the debate about whether Facebook is a public forum, the Court reiterated the US Supreme Court's wording in *Lloyd Corp. v. Tanner* —which held that property does not "lose its private character merely because the public is generally invited to use it for designated purposes'"— and came to the conclusion that "Facebook has not engaged in any functions exclusively reserved for the government".

This opinion was also shared in *Prager University* (2020), in which the **US** Court of Appeals for the Ninth Circuit dismissed claims against YouTube for allegedly censoring a video of "conservative viewpoints and perspectives on public issues". The case arose after YouTube classified some of the claimant's videos as "Restricted Content" which led to demonetization and age verification for their viewers. The difference with regard to the European assessment in similar cases is illustrated by the Court's emphasis that "the constitutional guarantee of free speech is a guarantee only against abridgment by government, federal or state". As the functions performed by YouTube are not traditionally and exclusively governmental, the platform did not transform into a state actor owing respect to fundamental rights. Also, the Court concluded that YouTube's public commitment to the right to freedom of expression must not be understood as an advertisement subject to consumer law, but rather "classic, non-actionable opinions or puffery".

In *Freedom Watch* (2020), the US Court of Appeals for the District of Columbia rejected, once again, the notion of big social media platforms as public forums or state actors. Moreover, the Court held that social media platforms can also not be considered "public accommodations"—which are obliged by the District of Columbia's Human Rights Act (DCHRA) not to discriminate—, as only physical spaces can constitute such a category. The claimant's allegation regarding the existence of a joint conspiracy of all important social media platforms —consisting of Google's demonetization of videos from conservatives (on YouTube), the exclusion of conservative leaning websites on Google's first page search results, routine suppression by Facebook's news curators of Republican-oriented news stories, and the "shadow-banning" of right-wing accounts by Twitter for political purposes— was rejected as lacking any factual basis.

c. Content moderation on an infrastructural level

The issue of content moderation does not only arise in relationships between users and platforms (consumer-to-business, C2B), but also on an infrastructural level (business-to-business, B2B). For example, providers of cloud services offer their services to other businesses on the condition that they do not use them for hosting particular content. This will become increasingly important in the foreseeable future, as for example the two main app stores (Google's Play Store for Android and Apple's App Store for iOS) exercise a fundamental gatekeeping function and have blocked apps in the past for not complying with their terms of use. While current court proceedings challenging such removals still focus mainly on an alleged abuse of market power and, thus, have focused on issues regarding competition law rather than freedom of expression, the decisions of infrastructure providers are at the same time increasingly understood as a form of content moderation.

One case which combines both topics is *Parler*. Parler —a well renowned social media platform amongst US conservatives—was suspended by Amazon Web Services (AWS) for lack of content moderation. The

platform challenged AWS's decision under competition law. The **US** District Court for the Western Court of Washington dismissed Parler's application for a preliminary injunction to prevent AWS's suspension of its services (2021). Parler was founded in 2018, describing itself as "a conservative microblogging alternative and competitor to Twitter." It used cloud computing services provided by AWS for hosting its application. During the 2020 presidential election, Parler saw an exponential increase in its online traffic as many users left Twitter to join the allegedly "uncensored" alternative. AWS noticed an increase of "illegal, harmful, or offensive" content posted by Parler's users, infringing AWS's Acceptable Use Policy. As Parler did not –in AWS's opinion– sufficiently moderate such content, it decided to suspend the service in the aftermath of the January 6 US Capitol attack. Similar to the *Freedom Watch* case, Parler's claim was based on competition law. It alleged a conspiracy of AWS and Twitter to reduce competition in the microblogging services market to the benefit of the latter. As AWS claimed that no such agreement with Twitter existed, and Parler was unable to provide any evidence in this regard, the case collapsed.

2. Claims to remove content or accounts

Social media platforms have not only been ordered to "carry" particular content, but also to remove content they are hosting. Many of these cases deal with the same two core issues: The first is the monitoring of content and the assessment of its illegality —which many deem fundamentally incompatible with the idea of an open internet where everybody can publish its ideas. The second is the question of the (extra-) territoriality of court orders to remove content. Should the illegality of content in one state lead to its global removal? Or does it suffice when the contested content is not accessible anymore from the court's home-state (geo-blocking)? The first is under constant suspicion of provoking a "race to the bottom", in which only the most harmless content, allowed in every single country of the world, can stay online. The latter seems like an unsatisfying result for victims who would like to have defamation and lies against them removed from the digital public sphere.

The first important decision by a Human Rights Court regarding the obligation to remove content (and a platform's liability when failing to do so) did not address the duties of the biggest social media platforms known worldwide. Rather the case concerned the comment section of an online national newspaper. In *Delfi As v. Estonia* (2015), the **European** Court of Human Rights' (ECtHR) Grand Chamber concluded that Estonia did not breach Article 10 of the European Convention on Human Rights (ECHR) —the right to freedom of expression— when it held Delfi AS, an online news outlet, liable for comments made by its readers. The Court considered that the award of damages was prescribed by Estonian law and served the legitimate aim of protecting the reputation and rights of others. Hence, the Court concluded that since Delfi set up the electronic infrastructure for allowing defamatory statements —and should have known that the specific article could have the potential to produce hundreds of angry, threatening comments—, it could be seen as a "publisher" or "discloser" of the comments. Thus, the award of damages (€320 in this case) met the threshold of being "necessary in a democratic society" and did not violate the ECHR. The Court emphasized that defamatory information posted on the Internet could potentially remain there indefinitely and cause much greater harm than comments in traditional print or broadcast media. The Court's decision was not undisputed: Judges Sajó and Tsotsoria considered the judgment, in their joint dissenting opinion, as "an invitation to self-censorship at its worst". They emphasized that states "by putting pressure and imposing liability on those who control the technological infrastructure […], create an environment in which collateral or private-party censorship is the inevitable result".

In 2023, the Grand Chamber of the ECtHR upheld in *Sanchez v. France* the Fifth Section's decision from 2021, which argued that the criminal conviction of a politician for comments published by other users on his Facebook page does not violate the claimant's right to freedom of expression. While the Court stressed the importance of protecting freedom of expression in the context of political debate, it found that the French courts' decision to convict the claimant had been based on relevant and sufficient reasons linked

to his lack of vigilance and responsiveness in monitoring his page for comments that could violate community standards or be unlawful. In an *obiter dictum*, the Court even stated that "a minimum degree of […] automatic filtering would be desirable" for the moderation of Facebook pages. Given the complex contextual assessment necessary to determine the legality of speech, such a call for Artificial-Intelligence-driven filter systems by a Human Rights Court is rather surprising.

However, the ECtHR has also acknowledged that a legal system does not necessarily, under all circumstances, have to hold a provider liable for defamatory comments published on its platform. The Court pointed out in *Tamiz* (2017) that the liability accepted in *Delfi* concerned a news portal run on a commercial basis, which are circumstances not comparable to a social media platform where the platform provider does not offer any content, and where the content provider may be a private person running a website or blog as a hobby. Taking into consideration the fact that information society service providers (ISSPs), such as blogging platforms, play an important role in facilitating access to information and debate on a myriad of political, social and cultural topics, the state's margin of appreciation is a wide one.

The question of a State's margin for appreciation also arose in the **European Union**: In *Glawischnig-Piesczek* (2019) the ECJ ruled that EU law, specifically the eCommerce Directive, does not preclude a Member State from ordering a social media platform to remove or block content that has been declared unlawful, or content that is identical or equivalent to such unlawful information. The monitoring of identical content or "equivalent content" qualifies as a form of specific monitoring: "monitoring in a specific case". Thus, it does not violate the prohibition of general monitoring obligations under EU law. Recognizing "equivalent content", for the ECJ, is not an unreasonable burden for hosting services, as long as it is not required to "carry out an independent assessment of that content". The Court also held that a removal order could apply globally, and left it to the Member States to determine the geographic scope of restrictions within the framework of the relevant national and international laws.

Accordingly, the **Austrian** Supreme Court ruled that Facebook must cease and desist from publishing all hate postings —verbatim re-postings or re-postings using words having an equivalent meaning—against Austria's Green party leader, Dr. Eva Glawischnig (2020), and delete them not just in Austria but worldwide. The Tribunal found that the cease and desist injunction imposed on Facebook was sufficiently specific in regard to the required conduct and did not impose a disproportionate or excessive duty on Facebook to monitor its content.

Similarly, the District Court of Frankfurt, **Germany** (2022), ordered Twitter to cease and desist from disseminating specific defamatory statements concerning the Anti-Semitism Commissioner of the state of Baden-Württemberg. The Court held that a social media platform must remove unlawful content as soon as it obtains knowledge of it, deriving this obligation from the platform's duty of care. This was also extended to equivalent statements which are substantially the same. Moreover, the platform must not only act upon notification, but also prevent the publication of predefined illegal content. While *general* monitoring obligations are precluded by European Union law, such *specific* monitoring obligations are possible (see *Glawischnig-Piesczek*). In the *Künast case* (2022), the District Court of Frankfurt awarded damages to Renate Künast, a German politician, for the continuous publication of misquotations on Facebook that, according to the Court, violated her rights. The Court also ordered Facebook to prevent the publication of new memes that include similar unlawful content by monitoring the users' uploads.

Before the *Glawischnig-Piesczek* case was decided by the ECJ, **German** courts had been more reluctant to issue removal orders, as demonstrated by the judgment of the District Court of Würzburg in *Anas M* (2017). There, the Court refused to grant an injunction forcing Facebook to take down a selfie showing the claimant, a Syrian refugee, with the German Chancellor Angela Merkel, falsely accusing him of taking part in several terrorist attacks in Berlin in 2016. It held that Facebook was not liable for illegal content on its platform, unless the content was reported and clearly illegal. Unlike the District Court of Frankfurt in later decisions, the Court concluded that service providers don't have a general duty to examine its users' con-

tent. Requiring filter mechanisms, the Court argued, would also affect lawful content and, thus, violate the users' right to freedom of expression. In response to a claimant's report, Facebook had geo-blocked the post in Austria and Germany. In the opinion of the Court, this measure sufficed to fulfill Facebook's obligations, as the Court has no extraterritorial power and cannot impose its laws outside Germany.

Similar cases and jurisprudence trends have also occurred in Latin America, particularly in Brazil and Argentina. The Court of Appeals of the State of Rio de Janeiro, **Brazil**, decided in *Lancellotti* (2016) that Facebook must remove posts containing false information about the actress Giovanna Lancellotti Roxo, suspend users' accounts targeting her, and pay damages for failing to protect her dignity. The Court applied consumer protection law to solve the case and found that Facebook's failure to address the problem of harmful and offensive fake Facebook profiles and communities violated the claimant's rights. Due to the personal distress, and damage to her professional life, caused by the content, she was considered to be entitled to financial compensation.

In *Vanucci* (2016), a Civil Federal Court in **Argentina** compelled Twitter to remove defamatory posts against the claimant, an Argentinian model, which violated her dignity. However, the Court also pointed out that this measure cannot be applied indiscriminately to all future content, but only to specific publications the claimant identifies and which, on the basis of a standard of proportionality and reasonableness, are found to violate her dignity. It underlined that Twitter cannot be required to carry out a prior examination of the content disseminated and hosted on the platform due to the amount of content uploaded. When assessing Twitter's liability for hateful comments posted on its platform, the Court also took into consideration its current efforts against hate speech by developing a joint Code of Conduct with other social media companies.

An interesting, and rather unusual, decision was made by the Civil Court of São Paulo, Brazil, in *Dória Junior* (2017), when it denied the claimant's request to remove a Facebook event that allegedly violated the peace and defamed the Mayor, João Dória, but ordered Facebook to provide the IP addresses of the users behind the content. The event, which was created to protest a controversial decision issued by the Mayor, and was located close to his house, attracted more than 25,000 people. The Court held that the content was a form of valid criticism towards a public figure, that there was no intention to defame the Mayor personally, and that the fear of disturbance near the claimant's house did not justify removing details of the event. However, Facebook was ordered to hand over the IP addresses of the users, as the Brazilian Constitution "does not allow anonymity".

The question of the geographical scope of a court's order was central in the *Ramdev* decision by the Delhi High Court, **India** (2019). The claimant, an Indian yoga guru, asked for the global removal of defamatory content from Facebook, Google, YouTube and Twitter. The platforms argued that they were ready to geo-block content in India, but that the global removal of content could lead to a conflict with laws from other jurisdictions, and, in particular, incentivize the practice of "forum shopping" by claimants who would choose the most restrictive of all possible jurisdictions to bring forwards claims to remove content. The Court undertook an extensive analysis of jurisprudence from other jurisdictions, such as the USA, the EU (referencing *Glawischnig-Piesczek*, see above), Canada, and Australia, and held that once defamatory content was uploaded from India and was made available globally, access to such content (once ordered by a competent court) should be disabled world-wide and not just restricted to India. In doing so, the Court differentiated between content uploaded from India, which must be removed globally, and content uploaded from outside of India, which must only be non-accessible from India.

In February 2020, the Delhi High Court in **India**, ordered Instagram and search engines to remove anonymous #MeToo allegations of sexual harassment against artist *Subodh Gupta*. The Court held that such allegations, which were published by an anonymous account (called "Herdsceneand"), "cannot be permitted to be made in the public domain/published without being backed by legal recourse". Hence it ordered Instagram to take down the posts and Google to remove links to search results containing information on

the allegations (for more jurisprudence on the responsibilities of search engines see below). An unusual detail of the case was the anonymity of the user behind the account, who was nevertheless allowed to submit its views (via its lawyer) to the Court without revealing their identity.

Just like in *Delfi* (ECtHR, see above) the liability of a news platform in relation to its reader's comments was at stake before an Australian court. In *Voller v. Nationwide News* (2019) the Supreme Court of New South Wales, **Australia**, held that media companies could be considered publishers of comments left by users on their public Facebook pages. Nationwide News and other news outlets shared links to their stories on their Facebook pages about the claimant, Dylan Voller, a juvenile who was mistreated at a detention facility. The articles provoked defamatory comments by other users. The Court, relying on jurisprudence from Australia and other Commonwealth countries, concluded that by maintaining Facebook pages, the media companies made it possible for these comments to become visible and harm the claimant. Thus, the defendants could be considered as publishers of the comments. Furthermore, the Court argued, the media companies should have been aware that stories on this issue would have led to defamatory comments, but failed to make such an assessment and to prevent them. The Court was particularly adamant that the media companies had the technical and resource capacities to moderate and even hide all comments before they became visible, but failed to do it. The Court also dismissed a freedom of expression argument brought forward by the news platform, finding that the media companies maintained Facebook pages for purely financial gain.

In **Ireland**, claimants have successfully sued Facebook for the "misuse of private information" in several cases, forcing it to remove content and to pay damages. In *J20* (2020), the High Court of Northern Ireland held that Facebook was responsible for its users' publications regarding information about a man's (J20) children, and the accusations that he was a police informant. Since the platform had received a notice on this content and did not act upon it, the claimant was entitled to damages of £3,000 for his emotional distress as a result of the postings. In *C.G.*, the Northern Ireland Court of Appeal found that the disclosure of a sex offender's photograph, name, address and previous offences, on a Facebook page violated this individual's right to privacy. The Court based its finding on the ECJ's *Google Spain* decision. The Court said that Facebook was —just like in *J20*—liable as it did not act upon notice. The Court clarified that Facebook was under no obligation to actively monitor the content it stores, and did not have to proactively remove the private information.

In the **United States** Section 230 has been used in cases in which social media platforms were tried for illegal comments as an ultimate shield of protection. The US Court of Appeals for the Ninth Circuit emphasized this protection in *Caraccioli v. Facebook* (2017), when it denied claims based on defamation, libel, intrusion upon seclusion, public disclosure of private facts, false light, and others. In this case, an unknown party created a fake Facebook profile of the claimant with multiple sexually explicit images and videos of him. Facebook initially denied the claimant's request to remove the user's account and only did so after the claimant's threat of legal action. The Court pointed out that Section 230 protects Facebook from all liability arising due to its position as "publisher or speaker" of another "information content provider". For the Court, only if Facebook itself could be considered to be the "information content provider", a claim could possibly succeed. Merely reviewing the content, however, did not suffice to qualify a social media platform as an "information content provider".

Up to now, the strongest challenge to Section 230 was the *Twitter v. Taamneh* case in 2023. Several social media companies were sued for their alleged failure to remove terrorist content from their sites by the families of victims of an Islamic State terrorist attack that took place in Istanbul in 2017. The social media platforms were accused of "aiding and abetting" a designated foreign terrorist organization in an "act of international terrorism" under the Antiterrorism Act. This brought up serious questions regarding the extent of the immunity provided for by Section 230, which were discussed heavily by the legal community. While the claimants were successful before the District Court and the US Court of Appeals for the Ninth Circuit, the Supreme Court reversed these decisions. The Court, however, decided to spare out all issues

concerning Section 230, and decided solely on the basis of the Antiterrorism Act. It emphasized that the platforms' relationship with ISIS "appears to have been the same as their relationship with their billion-plus other users: arm's length, passive, and largely indifferent", and that there is no proof that the social media companies were consciously trying to participate in the terror attack. Even when considering a duty on the platforms to remove terrorist content, "it would not transform [the platforms'] distant inaction into knowing and substantial assistance that could establish aiding and abetting '' under the Antiterrorism Act. On the same day, the Supreme Court also vacated the Ninth Circuit's decision in *Gonzalez* —in which Google was considered liable under the Antiterrorism Act, as YouTube's recommendation system was deemed partly responsible for the 2015 Paris terror attacks. SCOTUS ordered the Ninth Circuit to reconsider the complaint in light of the *Twitter* decision.

3. Claims for information (and the right to publish them)

Content moderation in a narrow sense focuses on the (non-)removal of content. However, there are also several cases in which individuals or organizations have requested information either on the individuals behind questionable posts by anonymous accounts, or on the system of content moderation set up by social media platforms.

Victims of defamation often do not only want content to be removed (and hold the social media platform accountable), but also want to pursue legal actions against the users who published the content in the first place. If they are anonymous, plaintiffs have requested, for example, information on the user, such as their email and IP addresses. While many legal systems do provide a legal basis for such claims (e.g., Section 21 of the **German** Telecommunications-Telemedia Data Protection Act), social media platforms in the **US** are once again shielded from such claims. In *Nunes* (2020), the Circuit Court in Virginia dismissed a lawsuit brought by US Congressman Devin Nunes. Nunes had invoked the state power to identify anonymous critics who had posted critical comments using the satirical Twitter pseudonyms "@Devin's Cow" and "@Devin's Mom". The Court pointed out that any lawsuits seeking to hold platforms like Twitter liable for exercising a publisher's traditional editorial functions (such as deciding whether to withdraw, publish, alter, or postpone content) were barred by Section 230.

Civil Society Organizations have lodged judicial actions that inquire, on an abstract level, about the system of moderation put in place by social media platforms. The Paris Court of Appeal, **France** (2022), held that Twitter had to provide information on their measures to fight online hate speech. Six French organizations approached the Court after their research indicated that Twitter only removed under 12% of the tweets that were reported to them. Furthermore, the plaintiffs also sought information on the resources Twitter dedicated to the fight against online racist, anti-Semitic, and homophobic speech, and incitement to gender-based violence and commission of crimes against humanity. The Court ruled that Twitter had to provide this information, as it was a necessary precondition to enable the organizations to determine whether to file an application under French law regarding Twitter's failure to remove, promptly and systematically, hate speech from their platform.

It is not always social media platforms which decide not to deliver information. In *Twitter v. Barr* (2020), Twitter challenged the **US** government's order to not publicly release its "Transparency Report", as it contained classified information. In its report, Twitter disclosed the amount of "national security legal process[es]" it received from the Foreign Intelligence Surveillance Court (FISC). In a case that underscored a clear rift between the First Amendment and national security concerns, the Court denied Twitter's motion and found that the classified declarations submitted by the Government satisfied, both substantively and procedurally, the strict scrutiny required to justify a content-based restriction and a prior restraint.

III. Cases against public officials and institutions

As shown above, social media platforms in the **US** are protected by Section 230 from any kind of claim based on the publisher's traditional editorial functions, such as deciding whether to publish, withdraw, postpone or alter content. However, deciding on such claims becomes substantively more complex when the opponents are not social media platforms, but state authorities operating a social media page in their official capacity.

In *Davison v. Rose* (2017), the **US** District Court for the Eastern District of Virginia dismissed a plaintiff's claims for violations of his First and Fourteenth Amendment rights against the defendants, who were members of a local school board. They banned the plaintiff from posting on their official Facebook pages, and even deleted comments that were overly critical of the defendants. The Court reasoned that the defendants were entitled to sovereign immunity in their official capacity and qualified immunity in their individual capacity. Since the law as to whether a Facebook page is a public forum is not yet clear, the defendants' actions did not meet the threshold required for violating a "clearly established" right of the claimant, which would be necessary for claims against them in their individual capacities. The District Court, however, found in *Davison v. Randall* (2017) that the Loudoun County School Board (LCSB) Chair's Phyllis Randall's Facebook page was a public forum and that she was not allowed to remove the claimant's comments and block him on her page. The decision was later confirmed by the Court of Appeals for the Fourth Circuit (2019). The school board's chair had argued that she was free in her decision to ban others "based on their views without triggering the First Amendment". However, the Court found there was sufficient evidence that the page was being used for official purposes, since it mentioned her position in the page's title, it had contact information relating to the county office and its official email address, and there was an explicit statement that the page's purpose was to be in touch with Loudon citizens. The Court held that the school board's chair exercised substantial control over the Facebook page —so that property questions didn't matter—, that the page "[bore] the hallmarks of a public forum"; and that it did not only contain "government speech". Thus, the interactive columns for public posts were a public forum. Hence, the moderation enacted by the chair, banning the claimant's allegations of corruption, amounted to prohibited "black-letter viewpoint discrimination."

A high-profile case concerning the same legal question can be found in *Knight First Amendment Institute v. Trump,* decided by the **US** Court of Appeal for the Second Circuit in July 2019. The Court found that the then-President's Twitter account was used for official purposes and that blocking critics was a government restriction, rejecting Trump's claim that his Twitter account was personal. The Court considered Twitter's interactive functions, such as replying, retweeting, and liking, to be forms of expressive conduct allowing individuals to communicate not only with the President but with thousands of others. It further established that the Twitter account was a public forum on the grounds that it was controlled by the government, and Twitter's interactive features made it "accessible to the public without limitation". Just like in *Davison v. Randall* (see above), the Court rejected the government's argument that the activity on the account was government speech, holding that Trump's individual tweets were, but the messages posted by users were private speech. Therefore, the Court concluded that then-President Trump violated the First Amendment when he blocked citizens for posting messages critical of him and his policies. Later, in March 2020, the Court denied a rehearing of the ruling.

However, the line between personal and official accounts is difficult to draw. In *Campbell v. Reisch* (2020), the **US** Court of Appeal for the Eighth Circuit decided that the mere election of a person does not "magically alter" the function of a social media account. Thus, Cheri Reisch, a Missouri State Representative, had not violated the claimant's First Amendment rights when she blocked the plaintiff from her Twitter account. The Court explicitly referred to the *Trump* judgment (see above) and pointed out that "not every social media account operated by a public official is a government account". Also, the reflection of the office a candidate is pursuing in the account name, or photos of them at work, do not suffice to turn an account into a governmental one. The Court distinguished the character of Reisch's account from the offi-

cial accounts in the *Davison* and *Trump* cases (see above) by noting that those accounts solely addressed governmental activity (such as the announcement of a governmental nominee or governmental response to a crisis). Reisch's account was closer to an election campaign newsletter. Consequently, her First Amendment rights "to craft her campaign materials necessarily trumps [the claimant's] desire to convey a message on her Twitter page that she does not wish to convey".

IV. Cases on state enforcement of private content moderation

The last two sections of this Special Collection deal with content moderation cases in a narrow sense. In these cases, individuals challenged either the social media platforms themselves, or government officials controlling a social media page, before a court to achieve the removal or reinstatement of content or accounts. However, not only individuals have tried to influence the moderation of content on social media platforms. Governments do too. Many national legislators have enacted new laws concerning the liability of platforms, with the explicit aim to "hold Big Tech accountable". One of the first and probably the most prominent example of such a law is Germany's "NetzDG", the "Network Enforcement Act". The name of the act reflects its main assumption: There are sufficient rules in liberal democracies on which content is legal and which is illegal, but these laws lack enforcement on private platforms. To increase the platforms' willingness to remove illegal content, fines for non-compliance were introduced. Enacted with the best intentions, it was used by authoritarian governments, such as Russia, Belarus, India and Malaysia, as a justification for similar legislations. In democratic states with a strong and independent judiciary, the possible negative effects of such laws, such as the "collateral censorship" Judges Sajó and Tsotsoria warned of in their dissenting opinion in *Delfi* (see above), can be diminished. Elsewhere, by contrast, where the system of checks and balances is weak, the effects are fully realized, as some of the following cases show.

1. Administrative proceedings to remove content

In the *Malaysiakini* decision (2021) by the Federal Court of **Malaysia**, a possible exemption from liability when acting upon notice was at stake. Malaysiakini is a Malaysian news portal which had published a press release issued by the Chief Justice. Subscribers had added critical and partly defamatory comments to the post. Even though the website removed the comments within twelve minutes after the police informed the website, the Court held the news portal guilty of contempt of court. It emphasized Malaysiakini's full responsibility for the use made by third parties of its own platform, as it controls who can post comments and it has installed filters to block certain words. Due to the controversial press release it republished, the Court considered that Malaysiakini should have known that this could attract illegal comments. In the opinion of the Court, liability exemptions for big social media platforms, such as Twitter—a "completely uncontrolled platform"—, did not apply to the news website, as its high degree of content moderation triggered its legal responsibility.

In **Russia**, the Tagansky District Court of Moscow decided over the years on several requests by Russia's Federal Service for Supervision of Communications, Information Technology and Mass Media (*Roskomnadzor*). In 2021, the Court imposed a fine of around USD$ 117.400 on Twitter for its failure to remove posts that called for participation in unauthorized rallies. In 2022, it was the Magistrates' Court №422 in the Tagansky District which held that Google repeatedly failed to filter its search results according to Russian law. Thus, the Court imposed a fine of roughly USD $52.800. The Court's order followed investigations by *Roskomnadzor*, monitoring whether search engine operators terminated access to those web pages that were subject to access restrictions in Russia. A few months later, the Tagansky District Court of Moscow held that also Meta had repeatedly failed to remove access to information that they had been instructed to delete, and imposed a fine of around USD $27 million. Meta had been repeatedly ordered by *Roskomnadzor*

to remove content it considered harmful to minors or inaccurate socially significant information —within 24 hours—from Facebook and Instagram. In the opinion of the Court, noncompliance with these orders justified a fine of 5 percent of the company's annual revenue. The Court also pointed out the "place of the offense" was where *Roskomnadzor* was located, hence questions regarding extra-territoriality did not matter. Also, it ruled that the law which served as legal basis for the fine was compatible with the Russian constitution.

In May 2021, the **Brazilian** Supreme Court ordered Twitter and Facebook, in an eye-catching case, to take down a number of accounts it had identified as spreading disinformation and threats directed against Supreme Court justices. In March 2019, the Brazilian Chief Justice Dias Toffoli initiated a criminal inquiry into insults against the Supreme Court. This came after months of growing criticism against the tribunal, as well as insults directed at its members, particularly by supporters of then President Jair Bolsonaro. As a result of the inquiry, a report was prepared and handed over to the Supreme Court. The Court found that the evidence demonstrated a "real possibility of the existence of a criminal conspiracy […] concerned with the dissemination of fake news; offensive attacks on individuals, to the authorities and to the institutions, among them the Federal Supreme Court, with patent content of hatred, subversion of order and incentive to breach institutional and democratic normality". Thus, the Court ordered, among other measures, Facebook, Twitter and Instagram to suspend the accounts of the individuals under investigation. After the Court found out that Twitter had only geo-blocked the accounts and their content—so that they weren't accessible from Brazil anymore but still could be displayed from other locations (or via VPNs even from Brazil)—, it imposed a fine for noncompliance to Twitter and directed the company to block content from the accounts, "irrespective of the means used to access the posts, or the IP [address] used, be it from Brazil or elsewhere".

2. Instruments of systemic cooperation

While some governments have initiated proceedings to force platforms to remove particular content, others have tried to enter into a dialogue with the social media platforms about their moderation mechanisms. The European Commission, for its part, has supported the tech industry in introducing several instruments, such as the "Code of conduct on countering illegal hate speech online" and the "Code of Practice on Disinformation". While they have been understood as mechanisms of "self-regulation", the involvement of the European Commission in its creation, and in particular the legal meaning they will obtain within the *Europe's Digital Services Act (see Article 45)*, qualifies them as instruments of "co-regulation", since they are a product of cooperation between private social media platforms and public governments.

The involvement of state governments in decisions about private content moderation was also at stake in *Adalah Legal Center for Arab Minority Rights v. Israel's Cyber Unit* (2021) by the **Israel** Supreme Court. The Supreme Court of Israel denied a petition regarding contesting the "voluntary enforcement" procedure of the State Attorney's Office's Cyber Department. This "voluntary enforcement" would begin with the Department taking notice of online publications that, *prima facie*, violated Israeli law. Then, the Department would refer the matter to Internet platform operators via a structured mechanism for reporting harmful publications. In turn, internet platform operators would need to address the report and decide, at their discretion, how to act and what to do under their community guidelines. The petitioners argued that the voluntary enforcement mechanism violated the separation of powers since the "last word" regarding a publication's lawfulness was in the hands of an administrative agency or an internet platform operator rather than a judicial court. The Court stressed that "the very possibility that the 'sword of coercive regulation,' which the government, or someone on its behalf can draw against the online platforms if their operators frequently fail to accede to the referrals is sufficient to show that we are concerned with a governmental act that requires some legislative authorization". In the absence of a specific legal basis for the Cyber Department's activity, its voluntary method could operate under the residual power granted to the government under Section 32 of the Israeli Basic Law, as long as its activities did not violate fundamental rights. Since

the petitioners were not able to prove that the "voluntary enforcement procedure" breached any fundamental rights or that social media platforms were not independent in their discretion, the Court considered the "voluntary enforcement" lawful.

3. Bans of services

Substantially more aggressive measures have been taken by less democratic states. Many have blocked internet services completely, such as social media platforms and the online encyclopedia Wikipedia, for not complying with their orders to remove content they deemed unlawful. In some instances, courts were able to stop such general bans, often after the platforms conceded to remove at least parts of the controversial content.

In May 2010, the Lahore High Court in **Pakistan** lifted a ban on Facebook, which was based on an individual's petition contesting Facebook's (and the Pakistani government's) lack of action against a "Everybody Draw Mohammed Day" page. However, while many legal experts stated that a blanket ban violated several rights enshrined within Pakistan's Constitution, including the right to freedom of expression and information, the Court based its decision on the assurance by the Pakistan Telecommunication Authority (PTA) "that the 'blasphemous material'" would no longer be available in Pakistan."

In **Russia**, the Tagansky District Court of Moscow granted in 2018 the request of *Roskomnadzor* to block access to Telegram in Russia because the company had not disclosed keys for decrypting messages sent over its network. Russia's Federal Security Services (FSS) had asked Telegram to submit messages and the respective encryption keys of six mobile numbers. Telegram declined the request as complying would, among other things, violate the right to privacy of its users. The Court reasoned that since Telegram operated in Russia it was under an obligation to comply with the laws of the country and to provide the Russian federal authorities with the means to decrypt messages sent over its network. The decisions on the means to block Telegram in Russia (an attempt which ultimately failed) was left to *Roskomnadzor*.

The Madras High Court in **India** reversed (2019) its own interim order, which directed authorities to ban the download of the TikTok application, after considering the safety features available on the app to deal with inappropriate and obscene content reported on its platform. The Court emphasized that TikTok had removed about six million videos containing "doubtful" content after its interim order. Thus, it was convinced that TikTok had a proactive take-down mechanism to deal with content abuse and complaints. However, the argument that such a ban also violated the platform's right to freedom of expression, as guaranteed by Article 19(1)(a) of the Indian Constitution, was dismissed because the rights of an intermediary such as TikTok, which created platforms for commercial purposes, are not protected.

In **Turkey**, the government issued a ban on all language editions of Wikipedia after it refused to remove two articles in English that claimed the Turkish Government was sponsoring terrorist organizations in Syria. The Turkish Constitutional Court held (2019) that the restriction—completely blocking all access to the Wikipedia website—was not justified by a pressing need and, as a blanket ban on access to the entire website, constituted a violation of freedom of expression and the right to access information. The Court emphasized that the interpretation of legal grounds such as "maintaining national security and public order" and the "prevention of offenses," in a broad sense, might lead to arbitrary practices and violate freedom of expression. However, the Court also noted the "goodwill" of independent and volunteer Wikipedia editors who extensively modified the encyclopedic entries and tried to reformulate them in a more impartial and objective manner.

Verdicts on the legality of blanket bans were not only issued by national courts. Regional courts like the Community Court of Justice of the Economic Community of **West African States** (ECOWAS) have

issued decisions too. It held in 2022 that the Nigerian government violated the petitioner's right to freedom of expression, and access to information and the media, by suspending Twitter in June 2021. The Nigerian authorities claimed the action was necessary to protect its sovereignty on the grounds that the platform was being used by a separatist leader to sow discord. The petitioners, however, claimed that the suspension was in retaliation for a flagged tweet by the Nigerian President, for violating Twitter's rules. The Court found that access to Twitter is a "derivative right" that is "complementary to the enjoyment of the right to freedom of expression". Because the Nigerian government was unable to show any legal basis for the suspension of Twitter, the block, in the Court's opinion, was in clear contravention of Article 9 of the African Charter on Human and Peoples' Rights and Article 19 of the International Covenant on Civil and Political Rights.

More information on internet censorship by disabling access to services, or even the internet as a whole, can be found in the *Special Collection of the Case Law on Freedom of Expression: Internet shutdowns in international law*.

4. Must-carry obligations

The above-mentioned measures all address the issue of an alleged failure of social media platforms to remove harmful content. They set up incentives, or legal obligations, to foster more active moderation or even censorship regarding content a government deems unlawful. In the **United States**, legislation with the opposite aim has been enacted in recent years. Conservative lawmakers are convinced that "Big Tech oligarchs" are "silencing" their voice on social media platforms. As a consequence of this assessment, the State of Florida, and the State of Texas, have passed "social media laws". The law of the State of Florida –*inter alia*– prevented social-media platforms from removing a candidate for public office from the platforms ("deplatforming"); limiting or prioritizing posts by or about political candidates; and censoring any "journalistic enterprise". It also required the platforms to apply "consistency" in their decisions to remove or limit posts or users; to allow users to "opt-out" of receiving a moderated feed; and to not change its conditions or standards more than once every 30 days. The law was challenged by two trade associations, *NetChoice* and the *Computer & Communications Industry*, representing a variety of internet and social media companies. The US Court of Appeals for the Eleventh Circuit upheld (2022) a preliminary injunction granted by a District Court against the regulation, as the majority of the contentious provisions were "substantially likely" to be unconstitutional. The Court stressed that social-media platforms engage in protected speech when moderating the content on their platform, and that, as private companies, they are entitled to curate a specific type of content and community for their platform.

The Texas law had very similar provisions, prohibiting "censorship" based on: (1) the viewpoint of the user or another person; (2) the viewpoint represented in the user's expression or another person's expression; or (3) a user's geographic location in this state or any part of this state. Just like in Florida, *NetChoice* challenged the law successfully before a District Court, which enjoined the enforcement of certain provisions of the bill. The Fifth Circuit Court of Appeals in Texas, however, granted a stay of the preliminary injunction (2022) on the grounds that content moderation did not constitute First-Amendment-protected speech and the bill was therefore constitutional. It referred to Section 230, stating that social media platforms "shall [not] be treated as the publisher or speaker" of other users, as an argument to further the idea that Congress didn't think the hosting of user content is a form of "speech", rather it is as mere conduit. The Court also emphasized that the common carrier doctrine "vests States with the power to impose nondiscrimination obligations on communication and transportation providers that hold themselves out to serve all members of the public without individualized bargaining", and that platforms were such providers. The Circuit Court's stay was later vacated by the Supreme Court in a close 5-4 vote, giving—once again—effect to the preliminary injunction issued by the District Court.


Global Freedom of Expression
Columbia University