Be Careful What You Wish For

Erik Tuchtfeld 2023-09-23T00:51:07

The European Court of Human Rights (ECtHR) is known as a fierce defender of human rights in the European Legal Space. It interprets the European Convention on Human Rights (ECHR), protects the human rights of 676 million citizens in 46 different states, including all Member States of the European Union, and even Luxembourg's European Court of Justice has to take its decisions into account when interpreting the Union's Charter of Fundamental Rights (Article 52 (3) CFR).

Lately, however, the ECtHR has issued some troubling statements on how it imagines content moderation in the digital realm. In May, the Court stated in *Sanchez* that "there can be little doubt that a minimum degree of subsequent moderation or automatic filtering would be desirable in order to identify clearly unlawful comments as quickly as possible" (para. 190) and has reiterated this position at the beginning of September in *Zöchling* (para. 13). This shows not only a surprising lack of knowledge on the controversial discussions surrounding the use of filter systems (in fact, there's quite a lot of doubt), but also an uncritical and alarming approach towards AI based decision-making in complex human issues. Also, the ECtHR's decision could be interpreted as constituting a positive obligation for states to order platforms to generally monitor their systems for unlawful content published by third parties. This would clash fundamentally with the legal framework established in the European Union for roughly a quarter-century, most recently confirmed by the enactment of the Digital Services Act (DSA).

To clarify one point from the beginning on: This is not a piece about the substantive nature of content. It does not concern the question, what *kind of content* should be allowed on the internet. Instead, it addresses the *issue of timing and automation*.

Platform Liability in the European Legal Space

To better understand the effects of the ECtHR's ruling, it helps to reiterate the principles of platform liability in the European Legal Space. While the US-American Section 230 shields service providers since 1996 from (nearly) all kind of liability for content published on their platforms, the European Union has from the very beginning on taken a slightly different approach. Since the year 2000, Article 14 of the e-Commerce Directive (from 17 February 2024 on: Article 6 DSA) stipulates that a hosting service provider is not liable for the content provided by someone else, on the condition that "the provider, upon obtaining such knowledge or awareness [of illegal activity or information], acts expeditiously to remove or to disable access to the information". This procedure is called "notice-and-take-down", as a hosting service provider is generally not liable, but once it is notified of illegal activities, it must act (and take the content in question down) to continuously enjoy immunity. This principle is further strengthened by Article 15 e-Commerce Directive (from 17

February 2024 on: Article 8 DSA), as it clarifies that there must not be “a general obligation on providers […] to monitor the information which they transmit or store”.

The importance of these provisions granting – in principle – immunity for service providers for the development of the internet as we know it today, can hardly be underestimated. An internet of content creators, be it by creating videos for YouTube, uploading images to Instagram, or posting controversial opinions on Twitter, is simply unthinkable if social media platforms would be legally responsible for all the pieces of content on their platforms, and thus had to check all of them before making them accessible to the general public. The text of Section 230 is even famously branded as the “26 words that created the internet”.

The ECtHR’s Stance on Content Moderation

The ECtHR was never very fond of the notice-and-take-down principle. In its first decision on platform liability, *Delfi AS*, in 2013 – confirmed by the Grand Chamber in 2015 – the Court held that national law imposing liability on a platform for third-party comments, even when they have a notice-and-take-down system in place, does not (necessarily) violate freedom of expression as guaranteed under the European Convention on Human Rights (ECHR). The Court emphasized that the States have a wide margin of appreciation when balancing the right to privacy (protected under Article 8 ECHR) and the right to freedom of expression (protected by Article 10 ECHR). The “Court would require strong reasons to substitute its view for that of the domestic courts” (para. 139). Thus, the Convention “may *entitle* Contracting States to impose liability on Internet news portals […] if they fail to take measures to remove clearly unlawful comments [such as hate speech and direct threats to the physical integrity of individuals] without delay, *even without notice* from the alleged victim or from third parties” (para. 159, emphasis added; confirmed in *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt,* para. 91).

Now, eight years later, the Court seems to be convinced that such a provision does not only fall within the States’ margin of appreciation, but that there is a positive obligation of the States to penalize platforms which do not remove illegal hate speech on their own initiative (without being notified). In *Zöchling*, an Austrian right-wing news portal published an article about the well-known Austrian journalist Christa Zöchling, the applicant in the ECtHR’s case. This sparked a series of comments by registered users, including death threats and massive insults. After the applicant asked the news portal to remove these comments, they were deleted within hours, the users were blocked, and the e-mail-addresses of the users were passed over to the applicant a couple of days later. The identification of the users, however, failed, as the e-mail-providers refused to share their names and postal addresses with her. The Vienna Court of Appeal decided that the platform had fulfilled its due diligence obligation by deleting the impugned comments immediately on the applicant’s request.

The ECtHR, however, was not satisfied with this. It criticized the lack of a notice-and-take-down system (which is rather odd, as the applicant notified the platform, and the platform took the comments down – it remains unclear, what the difference

to the system envisioned by the ECtHR would be). Also, – and this is the most problematic part – it reiterated the desire for a minimum degree of automatic filtering and noted that the Vienna Court of Appeal did not consider possible measures to be applied by the company to prevent defamatory content, such as a statement of the platform that unlawful comments were not only "undesirable" but prohibited. The Court emphasized that the platform could have anticipated further offenses, as past articles on the platform about the applicant also sparked offensive comments (all para. 13). Thus, the Court came to the conclusion that "the absence of any balancing of the competing interests at issue" violated the procedural obligations of the State under Article 8 ECHR.

Content Moderation as Human Rights Issue

Content moderation is an incredibly difficult task. Human communication is highly context-sensitive: What is used as an insult by some, can be a colloquial greeting between others. Emojis are regularly used as codes and develop a hidden meaning. Even the most abhorrent statements, such as death and rape threats, can – with little effort – be disguised so that they are only understood by those who know the context of a conversation. That is why thousands protested against the introduction of so-called "upload filters" by the EU's copyright reform in 2019, and even pioneers of the internet such as Sir Tim Berners-Lee have warned that such an obligation for automatic filtering would turn the Internet "into a tool for the automated surveillance and control of its users".

What the Court seems to overlook is that it's in particular the *ex post* nature of the notice-and-take-down procedure which is crucial to freedom of expression. Nobody's mouth is shut to prevent the expression of a statement, as repelling as it might be. But then, once the statement is out, when somebody becomes aware of it, deems it illegal and reports it to the platform (or the police), subsequent sanctioning can take place. This procedure is based on the assumption of individual freedom, with individual responsibility as its counterpart. It is the opposite of a police state which tries to prevent all kind of non-compliant behavior before it happens.

To draw an analogy to the offline world: A liberal state must not incarcerate citizens because their protests *might* exceed the threshold of legality, it should sanction them as a *reaction* to criminal acts which took actually place. Of course, this is not an absolute rule. There are actions which are so dangerous that they must be prevented in advance, in particular when the life of humans is in danger. In the offline world, the terrorist's bomb doesn't need to explode for the police to act, just like in the online world child sexual abuse material (CSAM) shouldn't be published before measures are taken against its dissemination. However, in particular when it comes to the exercise of political rights such as freedom of expression (or freedom of assembly), special care and restraint is needed, to avoid the establishment of a system of censorship. That's why, for example, the ECJ allows obligations for automatic filtering systems only regarding content which was declared to be illegal, so that the platform is not required to "carry out an independent assessment" of the legality (*Glawischnig-Piesczek*, paras. 45-46).

Sparks of Hope

While the ruling can be read as revolutionizing (probably by accident) the established system of platform liability in Europe, a more narrow interpretation remains possible. First, the ECtHR's desire for automatic filtering was stated (twice) as an *obiter dictum*, it didn't constitute a main pillar of the Court's reasoning. Second, the Court only found a violation of the *procedural obligations* of the right to privacy under the ECHR. A more sophisticated balancing of the rights in question by national courts might justify the same result, the Court's reasoning does not (yet) necessarily require the establishment of a general monitoring system. Third, up to now, the Court's two most important cases (*Delfi* and now *Zöchling*) on platform liability both concerned news portals. The Court explicitly stressed the responsibility of these platforms based on the controversies they spark with the articles they publish. Thus, one could argue that the ECtHR didn't introduce a general new framework for platform liability, but only sector-specific requirements for news portals. This would, however, stand in stark contrast to the EU's approach in the DSA. Here, the comment section of news portals serve as an example of an ancillary feature which should not trigger the legal obligations online platform services have to observe (recital 13, Article 3 (i)). Nevertheless, such a reading would help to contain the effects of the Court's ruling.

It remains to be seen if the ECtHR reconsiders its approach to platform liability and takes a more critical stance on automatic filtering mechanisms in future decisions. Its current position relies on private surveillance, facilitates systems of censorship, and harms the development of alternatives to the dominant social media platforms.
