Hogan Lovells

Using artificial intelligence to fight hate speech

07 February 2019

Hogan Lovells partner Winston Maxwell spoke at the executive roundtable on artificial intelligence and online hate speech, organised on January 31, 2019 by CERRE, the Centre on Regulation in Europe. The goal of the roundtable was to discuss what measures should be adopted to fight hate speech online and to look at the pros and cons of using machine-learning in that context.

The session started by a presentation by Michele Finck of her excellent paper presenting the advantages and risks associated with using AI to detect hate speech online. These risks are linked to the difficulty of defining and identifying hate speech as compared to nudity, for example, and the problem of automated tools being over-inclusive or under-inclusive. Harm to individual rights occurs in either case.

Winston reminded participants that balancing freedom of expression with the fight against harmful content has been around since at least 1997, when the U.S. Supreme Court decided *Reno v. ACLU.* More recently, the European Court of Justice has addressed various forms of content blocking mechanisms in its SABAM v. Scarlet Extended, Netlog, Telekabel and Google Spain/Costeja cases. The legal rule boils down to what we refer to in the EU as the proportionality test, which applies whenever a fundamental right such as freedom of expression is interfered with in the interest of protecting another right, such as human dignity, privacy or the right to property. The proportionality test involves three steps: First, the interference must pursue a legitimate objective — this will always be the case for fighting hate speech. Second, the interference must be provided for in a law that is clear and accessible. Third, the interference must be limited to what is necessary in a democratic society. Another way of expressing this third requirement is that you must look at a number of solutions to fight harmful content and choose the "least intrusive means" to achieve the objective. This third condition is where the choice of technological tools, including AI, comes into play.

When discussing measures to fight hate speech, proportionality applies at three levels. First proportionality applies at the stage of defining what content is illegal in the first place – the law must be specific as to what constitutes hate speech, and be adopted via democratic processes. Second, proportionality applies to choices on how to enforce the law. Who enforces it? The police? Private actors? And using what tools? Third, proportionality applies to the choice of mechanisms that ensure an effective remedy to an individual whose rights might be affected by a removal of content.

Michele Finck's paper points to the harms that can occur from AI mistakes: some illegal content will not be properly classified as such (false negatives), or some legal content will be wrongly classified as hate speech (false positives). But the lesson we've learned from EU case law on proportionality is that perfect tools are often the enemy of proportionality. We learned this from the SABAM v. Scarlet Extended case, for example, where an effective tool to protect copyright –deep packet inspection — was rejected in part because it threatened privacy rights. The proportionality test prefers a combination of imperfect tools – a kind of messy toolbox – which achieves a pretty good level protection without creating undue interference with other rights: costs for internet intermediaries, undue risks for freedom of expression or for privacy.

What kinds of tools do we find in this messy toolbox? The first tool, which we all take for granted is the notice and takedown mechanism itself. The notice and takedown mechanism in the European E-Commerce Directive and in similar US legislation already reflects a proportionality approach: to avoid a disproportionate chilling effect on freedom of expression, it is better to rely on ex-post notification and action, rather than on upstream filtering, even if this means that some harmful content will slip through. The notice and takedown tool remains the most important in use today. In addition, platforms are implementing technological tools, generally on a voluntary basis, to help fight illegal content. In specific cases, courts may also impose technical measures to block or remove content that has already been identified via the notice and action approach. These measures may involve outright removal of the content, or maybe just slowing down the propagation of the content if its illegal nature is not crystal clear. When in doubt, slowing down content may be more proportionate than removing it; the harm to freedom of expression will be lower. When choosing technical tools, another issue is the geographic scope. The definition of illegal content often depends on national culture, history and traditions. Limiting removal to a given geographic area might be more proportionate than removing the content globally.

For AI systems, key considerations will include whether to design AI systems so they err on the side of under-inclusion or over inclusion. If AI systems are used as a decision support system for human reviewers, how do we make sure that the human reviewers are not complacent and blindly follow the AI recommendation? The due process rights of the author of the content will need to be considered, including notification to the author, mechanisms for challenging the removal, and a swift appeal process. How tolerant should we be of anonymous speech? Should platforms require authors of content to identify themselves, and what level of verification is needed? The right to speak anonymously is protected as part of freedom of expression, but anonymity also facilitates hate speech, so a balancing act is required. Do we share data on sources of hate speech with the government and other platforms to help limit its spread?

Companies share data on cyberattacks to help limit propagation, so why not apply the same approach to hate speech? Yet data sharing would create significant privacy risks for individuals that need to be balanced.

The toolbox is full of choices that need to be combined in a way that assures a “pretty good”, as opposed to “perfect”, level of protection from hate speech while minimizing the interference with other fundamental rights. A one size fits all solution will not work, and this raises challenges for regulators. Regulatory frameworks are not used to dynamic, messy, imperfect, case-specific solutions, which is why the European Commission and other regulators try to give self- and co-regulatory solutions a chance to succeed before intervening with ‘command and control’ regulation. Detailed command and control regulation can become quickly obsolete, leading to disproportionate and/or ineffective outcomes. Experimental regulation is also important, because it helps find the best combination of imperfect tools to meet the proportionality test.

> Read the full article online