Regulating Online Hate Speech – Yuval Shany and Omri Abend

Although the problem of hate speech is not new, the extensive use of social media and other online platforms to disseminate hate speech raises particular law, policy and technology challenges.

- The speed and scale of dissemination generates almost uncontrollable ripple effects that may quickly transform words into action, as the recent massacres in Burma/Myanmar fostered by particularly vile hate speech on social media show.
- The 'market of ideas' rationale which justifies the toleration of offensive speech due to the high societal costs of censorship does not work well in social media environments controlled by algorithms that encourage that creation of self-contained eco-chambers.
- The global attributes of social media platforms create significant regulatory challenges given the different legal standards applicable in distinct countries, on the one hand, and the difficulty of foreign actors (content moderators and programmers) to fully understand local cultural, political and linguistic contexts, which could render apparently innocuous speech dangerous and vice versa.
- The limited role and capacity of states in monitoring online contents, situates social media companies and other internet intermediaries in the sensitive position of having to engage on content moderation in order to curb online hate speech.
- The power afforded to private entities to regulate speech and enforce community rules relating to hate speech, raises difficult due process and freedom of expression concerns, exacerbated by the limited transparency and accountability of many technology companies

It is against this background that our presentation will discuss the relevant *international law* standards governing the regulation of hate speech (article 19-20 of the International Covenant on Civil and Political Rights), acknowledging the particular challenges of distinguishing between legitimate and illegitimate contents (e.g., denial of historical events). The presentation then will discuss the policies employed by Facebook, focusing in particular on the interplay between machine and human decisions on content moderation (including content removal, quarantines, etc.), the human decision process and the interplay between decisions taken by Facebook and the domestic laws of the countries it operates in.

The presentation will then proceed to explore the potential for developing technological solutions to identify and handle online hate speech, discussing the challenges of flagging content and contexts through AI, and concerns about false negatives and positives. We will then discuss what are the policy alternatives which

the imperfect state of the technology recommends, and where academic research on online hate speech detection may contribute to this discussion.

- The research on online hate speech presented relates to an Israel Democracy Institute/Yad Vashem project involving as principal researchers Dr. Tehilla Shwartz Altshuler, Rotem Medzini, Prof. Karen Eltis, Dr. Ilia Siatitsa and Susan Benesch