On the Identification and Suppression of Hate Speech in Online Contexts

Johan Eddebo, Uppsala University–Centre for Multidisciplinary Research on Religion and Society, Sweden

The Asian Conference on Ethics, Religion & Philosophy 2023
Official Conference Proceedings

Abstract
This paper focuses the issue of identifying indirect hate speech on digital platforms. In previous studies, the authors have addressed automated flagging and suppression of hate speech in YouTube material. We found that such operations are characterized both by problems pertaining to the vagueness of the hate speech concept, as well as a problem compounding tendency of content creators towards circumventing identification and suppression efforts by mainly making use of indirect and tacit references. The indirect references chiefly function through various means of indicating levels of meaning above the immediate sentence, or the immediate signification, often by referring commonly held worldviews or ideological structures. This implies that automated suppression that relies on the flagging of keywords faces certain structural reliability issues in relation to these indirect communications, and important ethical and rights-related problems are also embedded here. For this reason, this paper will explore methods for the reliable identification of indirect hate speech. We will explore two methods: speech act theory and Grice's theory of incorporated cooperativity, and ascertain whether they separately or in combination can provide a framework for the reliable identification of indirect acts of hate speech. The paper also indirectly emphasizes the importance of worldviews, and the critical analysis of faith and worldviews, in regard to addressing contemporary political issues.

iafor
The International Academic Forum
www.iafor.org

Introduction

In contemporary practice within the digital environment, coherent and generally accepted definitions of hate speech for moderation and suppression, are generally lacking. Moreover, the definitions and templates in use are vague and to some extent arbitrary – something which invites a whole host of problems, especially when it comes to automatic moderation and censorship (Brown & Beall, 2008).

The Potential of Useful Definitions

In practice, however, hate speech or similar forms of disruptive discourses are not particularly difficult to define in principle. In the Roman Empire, for example, to publicly address someone in contravention of good morals was punishable by law. Similar sorts of injunctions against insults and defamations have been commonplace throughout the world's legal systems, and the very notion of speech acts so disruptive that they need to be suppressed is relatively straightforward (Sorial, 2015).

In other words, there are many ways we could in principle establish a coherent and non-vague definition of something like hate speech. One obvious approach is to establish a set of explicit moral principles that allow us to define which speech acts are to be prohibited. Another rather blunt way to go about this, is to simply from a consensus establish a set list of prohibited and clearly delineated speech acts, such as set of ideas which you are forbidden to express in public.

Three Ways of Ethical Foundationalism

But if we go with the first option, i.e. if we begin with a set of moral principles (which I here term ethical foundationalism), there are then basically three ways we can move forward and apply any moral principles to actual speech, three ways in which we can focus and ethically judge the character of the speech act, which all have somewhat different consequences in practice (Hietanen & Eddebo, 2022).

First of all, one can focus the consequences of the speech act and/or the ideas involved, formally speaking. If I make an utterance to someone, the effect in context is what renders the speech act permissible or not (according to one's chosen set of normative ethics).

Secondly, we can focus the teleology of the speech act, as exemplified in Aristotelian metaphysics and ethics. Given this perspective, it's the intentionality and tendency of the complete speech act that's in focus. The way in which the act is directed to certain ends, irrespectively of any actual consequences. So if one addresses a person with a slur, what renders the speech act impermissible or not is the intention in addition to the actual character of what is being said, and the way it will tend to impact (its teleology) upon a certain audience in a particular communicative setting. But since we here focus the teleology, i.e. the inherent tendency of the total act, it doesn't matter if we don't get any negative results. It is hate speech if it was intended as such, and/or if the speech act had the inherent tendency of hate speech, if it was the kind of statement that would have been received as hate speech, even if nobody hears the tree fall in the forest, so to speak.

Third, we have the purely formal approach. Here, the essential character of the ideas expressed or of the speech act performed is in focus. So what is then the difference in relation

to the other two positions? On this model, it is strictly the meaning of the ideas expressed that counts. It doesn't matter what effects the expressions have in practice, or what the inherent tendencies of the ideas or speech acts are – what's in focus is the objective assertion as such. If the meaning of what is being said is immoral or impermissible, it gets designated as hate speech.

This is similar to the notion of formal heresy of Catholic Christianity or the notion of *shirk* within Islam.

Nonetheless, in the context of our contemporary global digital framework that has a decidedly secularized character, and which spans a vast number of cultures, traditions and worldviews, there is a debilitating key issue with these approaches towards a comprehensive definition of hate speech based in the application of moral principles.

The Problem of No Common Moral Basis

The first significant problem is that such a definitional approach lacks a real foundation to build on. In contemporary secular society, there is no real agreement on what good morals are.

There is no common tradition able to supply unambiguous foundations of ethics to guide the automated mass censorship and moderation taking place on the digital platforms throughout the world. While many groups have strongly held values that in principle could play such a normative role, representatives of other traditions will not always agree – and the very notion of rigid, non-negotiable values often conflict with the overall ethical modus operandi of secular society (MacIntyre, 1988, 1990).

Still, to find common ground is not an insuperable obstacle, at least not between different religious traditions. There for instance seems to be plenty of room for a general agreement between, say, Catholics, Daoists, Hindus, Muslims and Buddhists, as to what constitutes good morals for interpersonal communication, and it is not inconceivable that we could establish an ongoing interreligious dialogue to support the moderation of communications in the digital sphere.

The Daost emphasis on harmony and balance is perfectly compatible with the Catholic principles of inherent human dignity as well as the Muslim's principle of the essential unity of the good of all creation inherent in the notion of tawhid; when one suffers, we all suffer.

Yet due to its vagueness and ambiguity, secular morals by themselves do not seem to be able to provide a lasting foundation for this sort of applied ethics, which possibly is one of the key reasons as to why the hate speech injunctions so far have been incoherent and arbitrary.

The emergence of a normative structure for global digital communications might for that reason possibly catalyze or at least motivate a return of religion to the public sphere. That said, the opposite tendency is arguably more probable – that an arbitrary set of secular ethics will become normative through these systems of social and narrative control, and that this process will tend to push aside the values and worldviews of religious traditions, not least the non-Western ones.

Problems of Automated Suppression

Apart from these basic problems of the anchoring of the values undergirding a coherent suppression of hate speech, actual moderation is beset by significant practical difficulties. One such difficulty boils down to the fact that people for the most part are pretty good at strategizing around these prohibitions and automated injunctions. Basic keyword filters are more or less worthless, and even more advanced algorithms trained by internet users to recognize and flag instances of hate speech are relatively easy to circumvent – for instance by the use of indirect speech or communication though evolving code words and symbols (Eddebo, Johansson, Hietanen, 2023).

So is an image of Pepe, the green frog of the American Alt-Right movement, stating "it's ok to be white" – is this an instance of hate speech? Of course not. And to suppress something like that would mean the end of liberal democracy as we know it. One can imagine what would happen to freedom of speech and the press if dominant platforms for communication began normalizing the suppression of content due to vague associations with politically undesirable ideas.

Nonetheless, through the contingent established associations of this particular cartoon and the slogan, its dissemination does indirectly support such discourses that, taken as a whole, fulfill the same function as immediate hate speech. This is still a problem.

Basic Principles for Capturing Indirect, Nested and Tacit Hate Speech via Algorithm

The ensuing question is then whether there could be more precise ways to automatically filter these complex and indirect acts of communication to minimize arbitrariness. Ways to mimic the complex awareness of indirect signification that human observers possess.

A human observer is immediately aware of complex discursive associations that sometimes latch onto otherwise innocuous content, which in itself does not motivate suppression - but sometimes can warrant caution with regard to the interpretation of the content matter and associated material. Could an AI in a similar sense be used to flag certain discursive clusters as warranting caution, and successfully identify potentially disruptive acts of communication of an indirect character? Could algorithms flag acts of communication that operate at a higher level of abstraction; through tacit associations, implied symbolic signaling or the like?

The short answer is yes, in theory. In practice, we are approaching this capacity, for instance through the recent rollout of the latest version of ChatGPT and similar forms of technology, which enable complex associations through absolutely massive layers of data. Given this background, there seem to be several models or theoretical frameworks that could profitably be used to enable an AI to accurately flag multi-layered discursive clusters.

Grice's model of cooperativity in dialogue is one example. His generic principles of cooperativity could simply be employed in a negative sense, i.e. to identify communication that's maximally uncooperative, and establishing this pattern as a proxy for discursive clusters that are potentially disruptive (Grandy, 1989).

The problem is that you need a comprehensive surveillance of the complete environment of communication. You need to collect the relevant set of discourses that populate the context, and you need some form of assessment of the communicative history of all of the agents

involved in the communicative situation to really be able to classify intended cooperativity. This is probably technically feasible at this moment in time, and the set of potential negative consequences is obviously immense.

One could also go with Searle's speech act theory, focusing communicative events that function as complex performative utterances. Searle's basic model is three-partite; we have the act of saying something and how that act is framed in a specific context; we have what you're more specifically doing in saying something, such as making a request or expressing gratitude, and we have the whole spectrum of implicit or explicit ways in which the speaker is trying to affect the audience (Tsohatzidis, 2007). So one in effect combines an analysis of formal content with rhetoric and performativity theory. To identify discursive clusters using a model such as Searle's could quite likely be done with some level of accuracy, but again, you need massive levels of data collection and a thorough surveillance of the agents involved and their respective communicative histories.

Conclusions

Notwithstanding these problems, it seems plausible structures of narrative control akin to the above-mentioned will be established unless the entire digital sphere collapses or something similarly catastrophic.

But one might ask if we couldn't perhaps get out ahead of this problem and nip it in the bud. If we couldn't scramble and rapidly foster the use of these technological approaches in some other way than outright or indirect censorship, if there are conceivable alternatives to flagging certain complexes of communicative acts as potentially disruptive, with all that this implies in terms of policy, the institutionalization of immense systems of control, and the indirect infringement of established rights? Is there something else we can do instead?

Back in the 90s, a group of researchers led by Hans & Laila Dybkjær attempted to employ Grice's theory of incorporated cooperativity towards something like this very end. The key idea was to establish a type of machine learning geared to expound upon the meta-communicative situation through strategically "asking the human user for clarification" (Bernsen Dybkjær & Dybkjær, 1996).

One way to employ this sort of AI in relation to moderation would be to introduce it as a party in digital communications online, in the sense of a Twitter or Facebook bot that enters into a group discussion and asks pointed questions of the purveyors of potentially disruptive discourses. So when your game loses the match, for instance, and you log on to your Facebook group to point out that the referees are morons, then this AI white knight uncannily reminiscent of a human user comes in and exclaims that: "X CAN POSSIBLY PERCEIVE THIS STATEMENT AS DEMEANING."

There might be many more options apart from this sort of strategy, but this sort of limited nudging approach seems possibly worse than mere censorship and surveillance; this is rather the explicit combination of mass surveillance with targeted and automated social engineering.

These seem to the options at hand. We have the potential end results of social engineering or outright censorship through mass surveillance to choose between in terms of automated hate speech suppression online.

And then we're suddenly also right back at the first problem once again. Because however we cut this, we're going to have to face the question of whose values and worldviews the automated hate speech suppression will reproduce.

In any event, I would argue that this entrenchment of values should be made very explicit. We should know exactly that the algorithms will tend to foster values ABC and worldviews XYZ. The entire process is much more likely, however, to be painted out as a neutral endeavor supporting vague, malleable and general conceptual constructs such as "human rights", while it in practice, and quite tacitly, will reproduce the values and worldviews that facilitate the bottom line of the corporate entities involved.

And not least for this reason, there's a sense in which this discussion and the institutional insufficiencies it reveals shows the continuing relevance of organized religion, even in connection to cutting-edge technological themes. In this particular situation, its relevance becomes clear in relation to fostering and guiding critical reflection around values and worldviews, and religion's potential to moderate undue excesses.

References

Bernsen, N. O., Dybkjær, H. and Dybkjær, I., (1996a). "Cooperativity in Human-Machine and Human-Human Spoken Dialogue," Discourse Processes, 21, 2, 213–236.

Brown-Sica M., Beall J. (2008). Library 2.0 and the problem of hate speech. *The Electronic Journal of Academic and Special Librarianship*, 9(2). https://digitalcommons.unl.edu/ejasljournal/99

Eddebo, J., Johansson, M., Hietanen, M. (2023). ”Automatic Identification of Hate Speech – A Case Study of YouTube Videos”. [Manuscript submitted for publication].

Grandy, R.E. (1989). “On Grice on language.” *The Journal of Philosophy* 86, 10, 514-25.

Hietanen, M., & Eddebo, J. (2022). Towards a Definition of Hate Speech—With a Focus on Online Contexts. *Journal of Communication Inquiry*, *0*(0). https://doi.org/10.1177/01968599221124309

MacIntyre, A. (1988). *Whose Justice? Which Rationality?* Notre Dame: University of Notre Dame Press.

MacIntyre, A. (1990.) “The Privatization of the Good: An Inaugural Lecture.” *Review of Politics* 52, 3: 344-61.

Sorial, S. (2015). Hate Speech and Distorted Communication: Rethinking the Limits of Incitement. *Law and Philosophy*, *34*(3), 299–324. http://www.jstor.org/stable/24572458

Tsohatzidis, S. L. (ed.), (2007). *John Searle's Philosophy of Language: Force, Meaning and Mind*, Cambridge University Press.