

Aristotelian approach to hateful speech moderation*

Maciej Kulik[†], Katarzyna Budzynska[†],
Medea Kfoczynska-Lukasz[†], Katarzyna Michalowska[†]

Introduction. Despite the fact that language abuse is troubling contemporary digital societies, the problem can be seen as early as in Ancient Athens, resulting in the death of Socrates and expulsion of Aristotle. Thus, the fact that the powerful tool to counteract is to be found in the Stagirite’s philosophy can hardly be seen as unexpected. The accuracy of the current hate speech definitions in detection of language abuse is limited due to the bias associated with the arbitrary selection of linguistic cues and protected characteristics. In this paper, we draw from Aristotle’s rhetoric and ethics to propose an alternative ‘hateful speech’ normative definition and verify it empirically to assess its accuracy in detection of harmful linguistic misbehaviour. The method, we use, is empirically-informed philosophical analysis. We follow the approach developed in (Koszowy et al., 2022) in terms of applying Aristotle’s philosophy to the empirically-grounded analysis of the online communication. First, we combine the relevant notions from Stagirite’s rhetoric and ethics and contemporary ethical accounts of hate speech in order to provide our normative hateful speech definition. Second, we verify empirically the accuracy of the proposed approach against 15 other accounts taken from linguistic, computational and ethical literature. Such an approach sets foundations for automated AI-based moderation.

Philosophical Analysis. We use Aristotelian analysis of love, hatred, justice and harm to solve problems encountered in contemporary discussions on hate speech and, at the same time, to avoid arbitrariness typical to the legal and computational accounts. According to Stagirite, love means “to wish what is good for his [the other’s] sake” (Aristotle, 1966, 1155b31-32, p. 194). On the other hand, the aim of hatred is not the good, but the evil of the other. In this respect, hatred differs from anger: “One who is angry might feel compassion in many cases, but one who hates, never; for the former wishes that the object of his anger should suffer in his turn, the latter, that he should perish” (Aristotle, 1926, 1382a30-32, p. 201). The possible co-occurrence of anger and compassion indicates that anger doesn’t aim at the evil of the other, but at restitution of justice that is violated by the illicit act. On the contrary, the aim of hatred is the annihilation of the other, that is, their ultimate evil. The contemporary accounts of hate speech touch upon notions such as dignity and harm (Herz and Molnar, 2012). Waldron (2014) combines both elements, as he identifies hate speech harms with denial of equal social standing to the individual who belongs to a certain group. Philosophical grounds for such an account can also be found in Aristotle: “Anger has always an individual as its object, for instance Callias or Socrates, whereas hatred [misis] applies to classes” (Aristotle, 1926, 1382a19-21, p. 201). However, rather than simply referring to the number, this specification amounts to distinguish the concrete individual and the abstract class. On Aristotle’s account, similarly

*We would like to acknowledge that the work reported in this paper has been supported in part by VW foundation (VolkswagenStiftung) under grant 98 542.

[†]Laboratory of The New Ethos, Faculty of Administration and Social Sciences, Warsaw University of Technology, maciej.kulik@pw.edu.pl

as in (Parekh, 2012), the act of hatred is inherently unfair, as it judges people on the basis of group membership instead of individual actions. Thus, the Aristotelian account of hatred, when applied to language misbehaviour, allows us to define hateful speech as a speech act that causes unfair harm by judging people on criteria irrelevant to the morality of their actions.

Empirical Verification. To verify our proposal empirically, we use a set of 30 utterances that represent verbal aggression, group libel, personal attacks and non-hateful yet emotional or judgemental expressions. Then, we ask two annotators to identify language abuse while using our proposal and 15 other accounts as annotation schemes to assess, how accurate these definitions are in detection of linguistic misbehaviour. The annotation of the set of 30 utterances, with our proposal and 15 other accounts as annotation schemes, allows us to determine the accuracy rate (percentage of the detected instances of hateful speech) alongside the error rate (percentage of the utterances mistakenly identified as hateful). This shows the accuracy of our proposal in terms of hateful speech detection, that sets foundations for the future moderation. Identification of the most reliable way to detect the harmful utterances sets foundations for automatic moderation, as it allows for development of AI-based Socratic virtual moderator in the sense of (Lara and Deckers, 2020), targeted at hateful speech moderation. This opens a chance for improvement, as the current moderation strategies rest upon manual interventions by human moderators and thus have only limited applicability.

References

- Aristotle. *The “Art” of Rhetoric*. William Heinemann, London - New York, 1926.
- Aristotle. *Nicomachean Ethics*. Oxford University Press, London, 1966.
- M. Herz and P. Molnar. *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Cambridge University Press, Cambridge, New York, 2012.
- M. Koszowy, K. Budzynska, and M. Pereira-Fariña. From Theory of Rhetoric to the Practice of Language Use: The Case of Appeals to Ethos Elements. *Argumentation*, 36:123–149, 2022.
- F. Lara and J. Deckers. Artificial Intelligence as a Socratic Assistant for Moral Enhancement. *Neuroethics*, 13:275–287, 2020.
- B. Parekh. Is There a Case for Banning Hate Speech? In M. Herz and P. Molnar, editors, *The Content and Context of Hate Speech: Rethinking Regulation and Response*, pages 37–56. Cambridge University Press, Cambridge, New York, 2012.
- J. Waldron. *The Harm in Hate Speech*. Harvard University Press, Cambridge, MA, 2014.