

Machine Ethos: the Goodwill, Virtue, and Wisdom of the Artificially Intelligent

Martin Hinton

University of Lodz

martin.hinton@uni.lodz.pl

As part of a broader project working towards the development of a system of rhetoric of machines, this study looks at how the Aristotelian concept of Ethos (Rhetoric II.1, 1378a6–20) relates to speakers who are not living beings and might usually, therefore, be thought to have no ethical standing. Some researchers are reluctant to recognise that interactions between humans and intelligent machines can be referred to as communication or that they can be considered as rhetorically interesting encounters. Others, however, recognising that such conversations will play an increasingly important role in our lives, have begun to consider how these interactions and their non-human participants can be described in terms of the rhetorical situations created.

While a great deal of attention has been paid to the related concept of ‘trust’ in AI, much less work has dealt directly with Aristotelian Ethos. One such study is that of Henrik Skaug Sætra (2024) and I shall consider his account along with my own, pointing out the similarities and points of divergence. Certainly, I concur with this aspect of his conclusion: ‘Issues such as trust, reliance, credibility, and more general issues related to machine acceptance and the relationship between humans and machines are all tightly linked to machine ethos’ (2024: 13).

Following the trend in the field, Sætra (2024) is also concerned greatly with issues of trust. He makes the importance of the relationship between ethos and trust clear: ‘evaluation of the machine’s intelligence, character, and goodwill will have significant consequences for whether the human trusts the machine’, but also, ‘ethos can thus be seen as a broader concept explaining trust’ (Sætra 2024: 6). An important distinction between ethos and trust is that ethos is emitted by the rhetor and trust is received in return. Sætra makes several interesting points about machine ethos, but his overall approach differs from mine in a fundamental way: although he argues that machine ethos will be mixed, he places particular importance on reputation or ‘prior ethos’ as opposed to the Aristotelian idea of ‘derived ethos’ (Braet 1992). This means paying great attention to the role of the creators of the machine and the image of the company producing it. My view is that such factors will be quickly forgotten in a longer term interaction, just as the reputation of a carmaker means little to someone who has owned a model for years, and the public image of a person does not survive intimacy.

The focus of my approach is on how machines come to have humanlike ethos, rather than on external non-human factors affecting our trust in them. Sætra is correct, however, to stress that we cannot dismiss the idea of a machine having ethos simply because it is a machine: ‘since ethos is about perceptions in the audience, the anthropomorphisation of machines, for example, could entail that machines are partly perceived as human – or alive – and susceptible to similar evaluations as humans’ (2024: 5). He also draws attention to the importance of the relationship built between user and machine and how the activity of the software may ‘generate the impression in the user that the machine is interested in them’ (2024: 10). This ability of AI to learn about the human and adapt itself to fit the user profile to seem ‘interested’ is both a great advantage in helping users feel appreciated and a great danger in exposing them to manipulation

(see Wilczynski et al. 2024). So, while I think his three part model of machine ethos featuring Makers, Machine, and Situated User, gives too much emphasis to what he calls the ‘origin and control ethos’, it remains an important contribution.

In my own model, I propose to divide the elements of machine ethos into three categories: those which have direct counterparts in human ethos, those which are in some way parasitic on factors of human ethos, and those which are machine specific. Into the first of these I place epistemological factors, such as access to the relevant facts and experience in dealing with them; into the second, humanlike, but on reflection clearly ‘artificial’ elements, such as accent and intonation, style and tone of language, and intimations of ‘character’; and into the third, intentions and competence of programmers and manufacturers, and issues relating to hardware reliability.

I discuss how each of these categories is relevant to the display and perception of the three qualities of Aristotelian ethos – goodwill, virtue, and wisdom – as well as considering the deeper question of what it actually means to ascribe such properties to a machine and how machine ethos will impact human-AI interactions and relations.

Aristotle. 2018. *Rhetoric*. Translated by C.D.C Reeve. Indianapolis: Hackett Publishing Company.

Braet, A.C. 1992. Ethos, Pathos and Logos in Aristotle's *Rhetoric*: A Re-Examination. *Argumentation*, 6: 307-320.

Sætra, H. S. 2024. A Machine's ethos? An inquiry into artificial ethos and trust. *Computers in Human Behavior*, 153, 108108.

Wilczyński, P., Mieleszczenko-Kowszewicz, W., and Biecek, P. 2024. Resistance Against Manipulative AI: key factors and possible actions. arXiv preprint arXiv:2404.14230.