Computers in Human Behavior 153 (2024) 108108

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

A Machine's ethos? An inquiry into artificial ethos and trust

Henrik Skaug Sætra

Department of Informatics, University of Oslo, Norway

ARTICLE INFO

Handling editor: Paul Kirschner

Keywords:
Ethos
Trust
Reliance
Rhetoric
Human-machine interaction
Human-robot-interaction

ABSTRACT

Every day we trust other individuals as we engage in social interactions in which various desirable outcomes depend on others acting the way we hope, or they have indicated. Trust extends beyond specific individuals, however, as we might trust unknown others – individuals, institutions, corporations, and governments. Some also say that we *trust* various artifacts, such as machines. But what is the basis of trust, and can we really trust *technology?* Trust is intimately connected to the notion *ethos* from the study of rhetoric and human persuasion, which is often used to describe various characteristics of the speaker, the audience, the relationship between the speaker and the audience, and the wider context in which communication and interaction occurs. In this article I explore to what degree *machines* can be considered to have ethos, and consequently whether ethos is a useful concept for understanding persuasive and credibility-related situations in HMI and by extension key aspects of human-machine *trust*. This allows us to draw on a long lineage of research from, for example, rhetoric, communication studies, and cognitive and social psychology to better understand the usefulness – or *not* – of using the notion of *trust* to describe our relationship with machines.

1. Introduction

Every day we trust other individuals as we engage in social interactions in which various desirable outcomes depend on others acting the way we hope, or they have indicated. Trust extends beyond specific individuals, however, as we might trust unknown others not to steal our bike when we quickly enter a store, that the banks where we place our money will give them back when we need them, and the government to provide us with protection against various ill fortunes (Amossy, 2001; Offerdal et al., 2021; Pilsch, 2018). Some also say that we *trust* various artifacts, such as mundane household equipment, our cars, and even that artificial intelligence (AI) and other complex computing systems can be worthy of our trust (High-Level Expert Group on Artificial Intelligence, 2019; Muir, 1994; Tavani, 2015). However, in various forms of human-computer interaction (HCI) and human-robot interaction (HRI) studies, the concept *trust* is regularly used without caveats or much problematization. But what is the basis trust, and can we really trust *technology?*

Trust is intimately connected to the notion *ethos* from the study of rhetoric and human persuasion (Brahnam, 2009; Weresh, 2012). The precise nature of ethos as a concept is debated, but it is often used to describe various factors related to the speaker, the audience, the relationship between the speaker and the audience, and the wider context in which communication and interaction occurs, and how such factors influence the credibility of an orator (Amossy, 2001). Ethos in rhetoric is regularly used to analyse how people persuade through a combination of the self-image they construct and how this is perceived by an audience (ethos) and appeals to logic and reason (logos) and emotions (pathos) (Aristotle, 2018). This use of ethos, which is the one adhered to in this article, differs from other uses of the concept in, for example, Hegel's philosophy, where it is used to describe the "ethical life" and the mores of a culture or particular society (Pinkard, 1986). The perceptions of an entity's credibility and trustworthiness will be established in part based on the norms and ethical conceptions of the society in which it is evaluated, but it is not the same. A machine's ethos is not it's ethics, but it's credibility as perceived by others.

In this article I explore to what degree *machines* can be considered to have ethos, and consequently whether ethos is a useful concept for understanding persuasive and credibility-related situations in HMI and by extension key aspects of human-machine *trust*. This allows us to draw on a long lineage of research from, for example, rhetoric, communication studies, and cognitive and social psychology to better understand the usefulness – or *not* – of using the notion of *trust* to describe our relationship with machines (Amossy, 2001; Giffin, 1967; Offerdal et al., 2021).

Furthermore, this exploration provides designers with concepts and a framework for developing new models for calibrating and monitoring the relationship between machines and their users, to be more or less

E-mail address: henrsae@ifi.uio.no.

https://doi.org/10.1016/j.chb.2023.108108
Received 7 August 2023; Received in revised form 9 November 2023; Accepted 21 December 2023
Available online 23 December 2023

persuasive and "trustworthy", and it gives regulators and other analysts a useful framework for analysing potential benefits and drawbacks of various types of machines and their interactions with humans.

Machine ethos is not, as mentioned, about the "ethics of machines". This lies in the domain of machine ethics (Allen et al., 2006; Anderson & Anderson, 2007, 2011), which deals with questions related to whether and how machines can *be* ethical, and deal with ethical challenges (Sætra & Danaher, 2022). However, machine ethos is tightly linked to a number of ethical implications related to, for example, anthropomorphism, deception, the gathering and processing of personal data, influence, and manipulation, which tend to be studied in AI and robot ethics (Sætra & Danaher, 2022).

In section 2, I I present a detailed account of ethos and relate this to research relevant to HMI trust. In section 3, this is used to develop a model of machine ethos, with a presentation of how machine ethos is built and shaped both before and during interaction. Section 4 considers the implication of the preceding sections and considers the relevance of machine ethos and how it might help designers and engineers deal more effectively and actively with the ethical and political implications of their creations. In this section I also present a number of objections to machine ethos, based both on the potential that ethos is not applicable to machines and that talk of machine ethos might give rise to a number of ethically problematic consequences.

2. Ethos and trust

Researchers in HMI have long been researching various aspects of the relationship between machines and humans, including issues of trust, reliability, credibility, persuasion (Muir, 1994; Tavani, 2015). It is also well documented that characteristics of machines, humans, and their relationship influence all these phenomena. These characteristics are also essential in the study of *rhetoric*, and particularly the study of *ethos* offers a long lineage of thinking useful for understanding how humans interact with machines.

In this section I first introduce ethos and emphasize various dimensions of the concept. Following this, I relate ethos to the main concepts used in HMI research to explore overlapping, but not identical, phenomena.

2.1. Ethos and rhetoric

> … ethos is a constantly renegotiated quality; it is an evaluation of the communicator that is performed by the audience and based on rhetorical artifacts, communicated at particular times as responses to particular problems (Offerdal et al., 2021).

Ethos is a fundamentally contested concept, and much time has been spent arguing over what ethos meant for the ancient philosophers, but also what it could and should mean for us today (Brahnam, 2009; Corts, 1968; Halloran, 1982; Sattler, 1947). As established in the introduction, I here refer to ethos as the evaluation of a communicator, and not the ethics of a culture or community, for example. Part of the controversy around the concept stems from the fact that it is at a "crossroads of disciplines" (Amossy, 2001), where quite different approaches found in, for example, sociology, psychology, and rhetoric, are used to analyse and approach the same concept. This has generated debates in which there is arguably a common core, but where different authors highlight and emphasize different aspects of the phenomenon. In addition to disciplinary differences, there are also several distinct methodological approaches to ethos in different disciplines, where psychology, and social psychology in particular, has generated much quantitative and experimental research and many *scales* for measuring ethos, for example (Andersen & Clevenger, 1963; McCroskey, 1966). Meanwhile, social theorists, rhetorical scholars, and philosophers have provided a plethora of qualitative accounts of ethos. In this article I draw upon both strands of research, while focusing specifically on the foundation of the concept as used in rhetoric, with the main aim of exploring to what degree the concept is useful in HMI.

Foundationally, ethos is one of three modes of persuasion: *Ethos* refers to the ethical appeal and persuasive power of a particular speaker in a given context – "arguments from authority" (Halloran, 1982) – whereas *logos* refers to the use of reason and logical appeal and *pathos* refers to the use of emotional appeal (Aristotle, 2018). While some see logos-based argumentation as the ideal form of argument (Braet, 1992), many argue that emotions and reason cannot be easily separated (Damasio, 2006), and also that the ethos of the speaker will inevitably have effects on persuasion. However, much is packed into seemingly straight-forward definitions of ethos, as we shall see.

Firstly, we must distinguish between *ethos* as a form of ethical proof relating to the persuasive power of a speaker or orator and *ethos* as a descriptor of how "we" – for example a profession or community – either *do* or *should do* things. Merton's "ethos of science" is an example of the latter (Merton, 1972; Segal & Richardson, 2003). Hegel also used *ethos* to describe the mores and "ethics" of a culture or society (Pinkard, 1986), or "living ethos" as the *sittlichkeit* or "public reason" of a community (Brudner, 2007). Such uses of ethos to describe to the ethics of various groups or professions are clearly distinct from my use of ethos, as I adhere to Aristoteles' use of the term to describe how the perception of orators affect their persuasive power (Aristotle, 2018). These conceptions are different both with regard to their focus on individuals vs. groups and to the use of ethos to describe some internal and "true" state of ethical convictions versus a perceived quality not *necessarily* reflecting a real or true state. However, the ethos I examine cannot be properly understood without also understanding the customs, laws, and moral attitudes of the community in which the orator is situated. This means that both conceptions of ethos are relevant, but only the former will be referred to as *ethos* in this article.

Secondly, when focusing on the former, it's important to note that ethos in rhetoric is a complex concept referring to quite different approaches to how the individual orator gains credibility and persuasive power through a) individual characteristics, b) their social situatedness, c) characteristics of their audience and the orators' ability to adapt appropriately to the context. This demonstrates the clear links between the norms and mores of a community and the perceived ethos of the speaker. While most researchers to some degree acknowledge all these facets of ethos, there is great variation in what is emphasised, and in the following I identify some of the main sources of both seeming and substantial disagreement over the use of the term. In addition, the distinction between introductory and derived ethos is established.

2.1.1. Individual characteristics and ethos residing in speech or text

The classical notion of ethos derived from Aristotle's *Rhetoric* (2018) is one where the orator – an individual – constructs and shapes a perception of themselves through speech (Amossy, 2001). While this approach to ethos emphasises perceived individual characteristics and speech, ethos is a dynamic concept liable to change through interaction (Andersen & Clevenger, 1963). This is so in part because an individual's characteristics might change, but even more so because ethos is about perceived and not real characteristics. Furthermore. Ethos is not only based on the words of a speaker, but also to their gestures, appearance, dress, and behaviour more broadly, which makes the concept potentially relevant also to people and other entities without language.

The three main components of an orator's ethos are *phronesis* (practical wisdom), *arete* (virtue), and *eunoia* (goodwill) (Aristotle, 2018, p. 55). These are in the modern literature often referred to as *intelligence* or *expertise*, *character*, and *goodwill* (Sattler, 1947). One classical representation of the main components of ethos and aspects of the different components are shown in Fig. 1, drawing on Sattler's (1947) account of ethos in ancient rhetoric.

Intelligence and character are mainly seen as the product of the orator's *virtues*, as evaluated and appreciated by the audience. Goodwill is generated by a perceived interest in the listeners *and* by a perceived

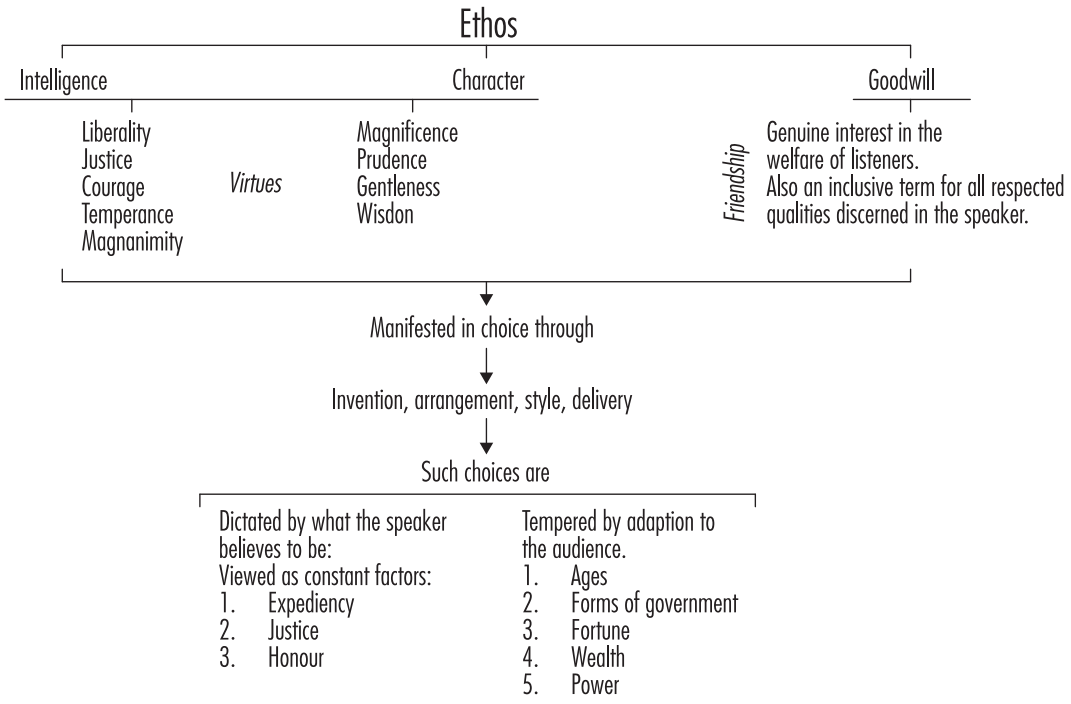

Fig. 1. Elements of ethos, based on Sattler (1947).

likeness between orator and audience, which is why Sattler (1947) refers to it as a term influenced by all respected qualities in the speaker. Moving downward in the figure, we see how ethos is manifested through choices related to invention, arrangement, style, and delivery, and these *canons* of rhetoric will be covered in section 3.3.1. In the bottom of the figure, we see how the orator's choices are based on their goals and principles and the *audience* characteristics, which is discussed further shortly.

2.1.2. Socially situated orators

The notion that the orator is a relatively isolated individual able to use their words to conjure up various conceptions of self not connected to reality is challenged by a range of contributions from, amongst other fields, sociology and feminist theory. For example, social theory and Bourdieu have been mobilized to demolish any Austinian notions that speech and the written word is decoupled from social situatedness and power relations (Amossy, 2001; Reynolds, 1993). This is also relevant for, for example, Hegel's notion of ethos as a social or communal form of ethics, and how moral consciousness and development takes place in families, cultures, and societies. Ethos, from this perspective, is a *social* act – one that cannot exist or be understood in isolation (Holiday, 2009; Reynolds, 1993). From this perspective, persuasiveness always comes from social positions and the access to and means to mobilize institutional and structural power. A speaker and "their discourse", is *authoritative* – persuasive – only when legitimated by a *situation* (Amossy, 2001). Moreover, ethos can be linked not only to situation, but also to *location* – "the sites on which an individual's social identity is constructed" (Reynolds, 1993). These points are often referred to as the postmodern critique, which states that ethos is not "in the text" – but in the interpretations, constructions, and projections of the audience upon the text and the orator (Brahnam, 2009; Holiday, 2009).

The postmodern critique does not entail that ethos cannot be analysed as the property of an individual speaker, however, as this does not entail some form of philosophical individualism or atomism. In this article I fully accept that the speaker is socially situated and use ethos only to describe how others perceive the individual. This relatively standard approach in rhetorical analysis requires that we analyse the individual's background, their social functions, and the norms and power structures in which the interaction between the orator and audience occur. This does not mean that "structure" absolutely trumps individual agency and characteristics, however, and Amossy (2001) convincingly shows how the individual and social perspective both can and must be combined to fully appreciate the persuasive power of an orator. However, it is also worth noting how some representatives of the postmodernist critique highlight how ethos can never exist in isolation, and that it contradicts the very idea that individuals have some intrinsic and consistent character (Holiday, 2009; LeFevre, 1987).

In the context of HMI, the postmodern critique is particularly useful for exploring to what extent a machine can be the kind of situated entity that becomes authorised through the social construction of their status as a) an entity b) situated in social contexts and institutions. This allows us to analyse the degree to which the machine is authorized by a group or community to act as a legitimate "spokesperson" able to act on others (Bourdieu, 1991). This perspective consequently necessitates considerations of the perceived moral and social status of machines (Gunkel, 2023), and also questions such as whether they can be *in* – or part of – social institutions (Sætra, 2023a). Such questions entail asking both whether the concept of ethos is applicable to machines *and* what the consequences of answering this in the affirmative would be. This will be further discussed in section 4.2.

Finally, this perspective of ethos raises question related to the ubiquitous nature of inequality in power and influence, as persuasiveness and all interactions are seen in light of the participants' positions and concomitant social capital and power. This stands in contrast to the liberal ideal in which all individuals are considered free and equal and in which free and rational exchange of opinions enables the discovery of the stronger arguments freed from any unjust preconditions (Sætra & Ese, 2023). This will be considered further in section 4.1.3, as such a perspective necessitates engagement with and awareness of structures of inequality and oppression, which means that developers and designers must at times choose whether to adhere to and exploit such knowledge or to actively break with and counteract power inequalities in human relations.

2.1.3. Audience characteristics and the orator's adaptive capacity

> … all argumentation depends, for its premises, as indeed for all its unfolding, on what is accepted, on what is recognised as true, as normal, as believable, as valid: through that it becomes anchored in what is social, the characterization of which will depend on the nature of the audience (Perelman (1989, p. 362), quoted in and translated by (Amossy, 2001, p. 5)).

Common to all approaches to ethos is the idea that there can *be* no ethos unless there is an audience – one, several or many. Speech and other acts will clearly potentially be both based on an individual or social ethics, and can have ethical consequences, but this is not the ethos I speak of here. The audience matters for understanding a speakers' ethos for two major reasons: a) their backgrounds, preferences, personalities, etc. make them evaluate speakers quite differently, and b) the effective use of ethos in part depends on successfully making the audience identify with the speaker, for example through awareness of and active use of *shared beliefs* and *knowledge* (Amossy, 2001; Weresh, 2012). Beyond beliefs and knowledge, it is also highly beneficial to generate the appearance of *preferences* being aligned, as this engenders audience trust and makes the speaker appear "good (*agathos*) and benevolent (*eunous*)" (Braet, 1992). The ability to forge a *bond* – also referred to as identification – happens through either natural likeness or the orator's ability to adapt to the audience (McCormack, 2014). Emphasis on the audience, then, makes us focus on source-*relational* attributes (familiarity, similarity, etc) and not only the source-*characteristic* attributes emphasised in 2.1.1 (Weresh, 2012).

However, it will at times be beneficial for the speaker not to appear to be *like* the audience, but to appear to be of a type the audience identifies with a particular set of qualities. For example, a very lively and expressive audience might find a seemingly careful, constrained, and modest speaker to be highly credible, particularly if the speaker seeks the position of their auditor or financial manager, for example. The point is not always to appear to be identical to the audience, then, but to know and adhere to their expectations for the role the speaker seeks to occupy. Being effective in the interaction with the audience requires the effective use of stereotypes and received opinion (doxa), which cannot be done without a proper understanding of who the audience are and what characterises them. It also highlights the need for an orator *not* to be wholly unique and novel (Halloran, 1982), but to always link the discourse to something already established and known – shared representations (Amossy, 2001).

Deeply ingrained in the concept of ethos is the idea that audience perception of the *same* speaker and interaction will vary based on several variables. Aristotle himself emphasised this, encouraging the orator to, for example, adapt to the audience's age (Aristotle, 2018). Empirical research has also long emphasised the need to focus on the effects on ethos of variables such as sex, occupation, educational, status, and political ideology (Andersen & Clevenger, 1963). It is also worth noting that an audience familiar with rhetoric will perceive more subtle differences in ethos than those unfamiliar with it (Andersen & Clevenger, 1963).

While these individual factors are important, the cultural context is also of crucial importance for ethos. Rhetorical ethos is for example used to show how an orator needs to communicate and present themselves quite differently to a North American or European audience and, for example, a Chinese audience (Campbell, 1998). This again demonstrates the link between an orator's ethos and the culture and norms in a specific context. As noted by Halloran (1982, p. 60), to "have *ethos* is to manifest the virtues most valued by the culture to and for which one speaks." In Athens, these virtues were, amongst others, justice, courage, temperance, liberality, prudence, and wisdom (Aristotle, 2018). Focusing on culture before the individual audience member is also the better approach for ensuring that one accounts for the foundational level of moral perceptions and attitudes that can be assumed to influence the individuals. As I'll discuss later, focusing on group characteristics will also have a range of benefits related to the reduced need to gather personal data to profile individual audience members.

After understanding the culture of the audience, more specific individual characteristics can be considered. However, it is important to consider the effects of the individual variables mentioned above as liable to change between cultures, and not to assume that, for example, *gender* or *age* has the same impact everywhere. Culture and the individual are always connected.

This perspective raises the question of how machines can adapt most effectively to their human audience in interaction. How they *can* is one thing, but if one concludes that they *can* it is crucial to then proceed to consider whether and how they *should* adapt, as I'll return to in more detail in the discussion of the ethics of machine ethos in 4.2. It also relates to the status of machines as social entities and how they are perceived and evaluated by different humans, as discussed further in section 3.

2.1.4. Introductory and derived ethos

Ethos is also often separated into introductory or prior (pre-discursive) and derived (built through discourse) ethos. While the orator *shapes* and *builds* ethos through interactions, in many cases the audience will have a *preconception* about the author, by having, for example, seen or heard of the orator beforehand (Halloran, 1982). Aristotle himself insisted that ethos must come from the *speech* itself and not preconceptions (Braet, 1992), but modern approaches to ethos recognize the importance of also accounting for audience exposure to relevant stimuli pre-interaction. This stimulus involves not just speech, but also various aspects related to the machine's history of *behaviour* and its general reputation. This means that effectively using ethos does not only rely on adapting the interaction to the audience, but also knowing what the "baseline" ethos is, meaning what perception the audience has of the orator before they start interacting (Amossy, 2001).

Some question the notion that ethos can be disconnected from the "true" nature of the speaker, and detest talk of "manipulating" or "constructing" ethos (Reynolds, 1993), even if it has always been seen by some as a way of, for example, *manipulating* judges through the strategic use of ethos, pathos, and logos (McCormack, 2014). Distinguishing between the perceived ethos and the "real" character of a machine is particularly tricky, as we'll return to in the discussion of objection to any talk of machine ethos at all in section 4.2. A pragmatic approach that sees ethos as at least partially detached from real characteristics or, for example, the ethical conviction of the orator, is here adopted.

It's relatively uncontroversial to consider rhetorical ethos as shaped and built through discourse and interaction, and that it is also heavily affected by conscious choices made by the orator, even if, for example, an orator's voice might at times "reveal" certain characteristics, which supports the "truth-will-out" theory suggesting that ethos cannot be faked (Andersen & Clevenger, 1963). It seems pertinent to assume that derived ethos stems both from uncontrollable and subconscious behaviour *and* the voluntary and goal-directed choices of the orator, and derived "ethos thus produced seeks to procure for the speaker a long-term benefit which could well make a difference" (Amossy, 2001, p. 21).

Ethos, then, is here considered to be about *appearing* credible and trustworthy, and this opens the door to dissimulation and deception (Brahnam, 2009). Approaching ethos *strategically* is consequently difficult to avoid, and this relates the exploration of ethos to other forms of strategic impression management and, for example, the use of intentional and strategic failure in HRI (Sætra, 2023c).

2.1.5. Ethos for non-individuals?

A key question for this article is whether non-humans *can* have ethos at all, and the degree to which human involvement will always be the main – or even only – determininant of the potential ethos of non-humans such as machines or corporations. Brahnam (2009), for

example, argues that it is questionable to consider ethos as something not connected to *humans*. However, 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. Anthropomorphism is a concept used to describe how humans project human-like qualities, and for example *intentions*, to non-human entities. That humans anthropomorphise machines of various kinds is well established, and the ethics of, for example, social robots, involves considerations related to how to balance the potential positive effects it has on HRI against the negative consequences related to, for, example deception caused by machines that encourage anthropomorphisation (Duffy, 2003). Regardless of one's position on the ethics of anthropomorphism, the fact that humans *do* anthropomorphise means that they are also likely to perceive a form of *ethos* in various machines – particularly the human- and life-like ones.

We also know that ethos *is* quite often used to describe non-individuals – such as government institutions and private corporations (Pilsch, 2018). This relates both to the idea of having a company ethos – or "how we do things" – and, more relevant for this article, the analysis of a company's ethos in corporate communication, for example (Isaksson & Flyvholm Jørgensen, 2010). Businesses are naturally very conscious about their ethos, and consider the use of style and its effects on how the corporation is perceived (Kallendorf & Kallendorf, 1985; Offerdal et al., 2021). While organizations are clearly composed of humans, their nature is quite distinct from that of individuals. Organisational ethos then, refers to some articulation of its personality, character, and individuality, and it is used both for external marketing purposes and for building internal cohesion (Pilsch, 2018). This strategic use of ethos in communication is fully in line with classical rhetoric, even if the ethos here belongs to a non-individual, and it shows how ethos is often *compound* – made by and sourced from several individuals and even organizations at once – like machine ethos, as I'll shortly return to.

In the examination of machine ethos, this non-individual kind of ethos will be essential for understanding how the ethos of the corporations and organizations responsbile for and otherwise linked to the machine influences the user's perception of the machine and thus the machine's ethos.

2.1.6. Summary

The preceding considerations have shown that ethos is a complex concept, and that there are multiple and quite different – but complementary – approaches to the concept. A key insight for the following sections is the idea that ethos is always something generated in *interaction* between a situated orator and audience *in* a specific social context (Andersen & Clevenger, 1963; Sellars, 2006). As the context changes, so does the fundamental nature of the interaction, and what is conducive and detrimental to ethos formation might radically change as well. The individual ethos is consequently tightly linked to the culture in which discourse takes place.

Furthermore, we can note that ethos is not something intrinsic in the speaker *or* the audience, I do not posit a direct and necessary correspondence between perception (ethos) and reality (the ontological or *actual* properties related to intelligence, character, and goodwill). Ethos is in a sense a social construct generated through interaction and consists of the *image the audience holds of the orator*. The notion of social construction and interaction is crucial for understanding ethos. In fact, ethos can be seen as the result of a series of processes in which the orator *constructs* an image of the audience in their mind and adapts to this image – not reality – while the audience does the same of the orator. In the words of Amossy (2001, p. 6), "the discursive construction of ethos is realized through a series of mirror reflections". A common assumption in rhetoric is that the "audience enjoys listening to speeches which mirror its own nature" (Braet, 1992, p. 313), and adapting to the audience is thus imperative unless other goals than maximum persuasive power is sought.

With these considerations we are almost ready to tackle questions related to machines' abilities to have intelligence, character, and wisdom, to use style and arrangement effectively, and to create bonds and relationships with the audience as social entities of some kind. First, however, I will connect ethos to trust and other overlapping concepts often used in HMI research.

2.2. Trust and ethos

> This concept of ethos appears to denote the degree of trust a listener is willing to place in the message of a speaker: It reflects a willingness to rely upon or show confidence in the speaker and his message (Giffin, 1967, p. 106).

The main goal of this article is not a philosophical exegesis of rhetorical scripture, but to show how the concept can be useful for understanding *trust* in HMI. Why, one might ask, use ethos, when so much research exists on HMI trust, reliance, and credibility? I'll argue that ethos provides a comprehensive theoretical framework for *uniting* many existing approaches, and that it is highly valuable to explore the relational and qualitative questions that emerge with the considerations of ethos, as opposed to the overwhelmingly quantitative and experimental approach to HMI trust often found in previous studies. The latter methodological approach has also led to a situation in which *trust* is operationalized and flattened in ways that removes it a great distance from how humans normally use and understand the term as applied to their relations with others.

2.2.1. Trust, reliability, and source credibility

> Trust of a speaker by a listener, called "ethos" by Aristotle and "source credibility" by Hovland et al. (1953) (Giffin, 1967, p. 106).

Much of the literature relevant to understanding ethos does not even use the term. This is because many disciplines are concerned with various aspects related to trust, source credibility, prestige, etc., and approach these phenomena which overlap with *ethos* using different terminologies and concepts (Andersen & Clevenger, 1963). Nevertheless, this research is often relevant for understanding aspects of ethos, and by briefly exploring the overlap of concepts, the usefulness of using ethos in HMI is further established, as concepts such as trust and credibility are often used in HMI research already (Glikson & Woolley, 2020; Muir, 1994; Sica & Sætra, 2023; Tavani, 2015).

In the media arts, one often speaks of *believability*, and Brahnam (2009) shows how this research can easily be linked to or conceived of as inquiries into ethos. The notion of a suspension of disbelief is relevant for understanding human-human interaction (HHI), but perhaps even more important for understanding the effects of anthropomorphisation and robot deception (Brahnam, 2009; Sætra, 2021c). Of key importance for Brahnam (2009) is how machine perceptions will be subject to a *tension* between our tendency to anthropomorphise machines – often referred to as the "Eliza effect" (Weizenbaum, 1976) – and what she refers to as the "Weizenbaum effect" – which describes the tendency to combat and mitigate anthropomorphisation (Brahnam, 2009). Credibility is another concept tightly linked to ethos, as persuasion through "appearing to be or being a credible person" is what ethos is all about (Kallendorf & Kallendorf, 1985). Our understanding and analysis of machine credibility and "believable agents" (Brahnam, 2009), for example, could be helped by the use of machine ethos as an analytical lens. Machine credibility is often explored through factors relatable to ethos, such as perceived competence, character, composure, dynamism, and sociability (Burgoon et al., 2000), and these are shaped by the machine's actions, including speech or text for machines designed to interact through words.

Source credibility is another widely used concept related to ethos (Giffin, 1967). Hovland et al. (1953, p. 21) defined it as "(1) the extent to which a communicator is perceived to be a source of valid assertions (his

'expertness') and (2) the degree of confidence in the communicator's intent to communicate the assertions he considers most valid (his 'trustworthiness')." It is consequently "essentially similar" to the concept of ethos, and also clearly related to *trust* (Giffin, 1967). This serves to show how similar phenomena have been studied under different labels, and also shows why it could be useful to unite some of the findings from these research strands (Pornpitakpan, 2004), for example through analyses of machine ethos.

The literature on ethos sees "creating trust, or a greater perceived similarity, familiarity, or liking between" orator and audience, as crucial for persuasion, and consequently incorporates the trust concept in the analysis of ethos (McCormack, 2014). However, the goal of designers and engineers will – or should – not be to *maximize* the credibility and positive ethos of the machine, as this will potentially involve excessively promoting anthropomorphism, with the risk of significant deception issues. In addition, it could lead to poor calibration between the machine's capabilities and the users perception of the machine, increasing the dangers related to, for example, *overtrust* issues where users excessively rely on machines in situations where this might be detrimental to them (Aroyo et al., 2021).

Some highlight how trust can be established through "showing similarity, creating a bond, and by maintaining good will", and stress how trust and credibility are also undermined with poor perceived character and goodwill (Weresh, 2012, p. 235). Others use ethos to analyse how public institutions can build appropriate trust in the public, and states that ethos provides a "clear foundation for the conceptualization of trust" (Offerdal et al., 2021, p. 248). What *is* trust, then? Giffin (1967, p. 104) long ago lamented the status of knowledge of the concept:

> The word "trust" has been prominent in our vocabulary for years; however, the concept is somewhat similar to Mark Twain's notion of the weather: Everybody knows about trust, but few people have studied it.

The concept of trust is prevalent in HMI research, and efforts to measure trust, understand "trust calibration", overtrust, etc., abound (Aroyo et al., 2021; Lee & See, 2004; Muir, 1994). A full account of this literature is beyond the scope of this article, and my main point is to show the obvious link between *perceived trust* and the development and shaping of *machine ethos*. Machine ethos describes a human psychological pheonomenon related to the evaluation of machines, and this evaluation of the machine's intelligence, character, and goodwill will have significant consequences for whether the human *trusts* the machine as well.

The similarities between the literature on HMI trust and my undertaking are many, and particularly the question of whether the *human* concept of trust applies to *machines* has been asked and attempted answered before (Sica & Sætra, 2023; Tavani, 2015). Tavani (2015), for example, answers in the affirmative, and provides an account of the concept of trust and how different machines are capable of different *levels* of trust. The question is always whether a human (trustor) can trust the machine (trustee). As with ethos, research on trust is also characterised by some disagreement regarding the appropriate perspective to take, such as seeing trust as an individual feature, an expectation, as acceptance of and exposure to risk, or as an institutional phenomenon (Beldad et al., 2010).

For our current purposes, we can accept the plurality of definitions of trust, but we require some common core to relate ethos to trust. One proposal of a description of a "trusting person" is provided by Giffin (1967, p. 104), who argues that it requires a) one person *relying* on something, b) this thing being an object, event, or person, c) that something is *risked*, d) that the trustor has a goal, e) that the achievement of the goal is not *certain*, and that f) the trustor "has some degree of confidence in the" trustee. The complexity of the concept is thus made evident.

2.2.2. Linking ethos and trust

Much effort related to HMI trust is aimed at finding definitions that allow machines to enter trust-relationships, often in ways that allow for the quantitative and experimental handling of the concept. One danger of such attempts is that the concept is impoverished and twisted, giving rise to very low definitional validity (Sætra, 2023b). To take the concept one step "back" towards the more traditional use of "trust" in natural language, which also helps show the relevance of *ethos* for examining trust properly understood, I point to an influential distinction between two different *types* of trust: *benevolent*-based and *integrity*-based trust (Levine & Schweitzer, 2015).

As should be immediately evident, these conceptualisations of trust are highly relevant to the analysis of ethos. Integrity relates tightly to the character of the orator, and entails that "the belief that the trustee adheres to a set of acceptable ethical principles, such as honesty and truthfulness" (Levine & Schweitzer, 2015). Benevolence is tightly linked to the goodwill discussed in the literature on ethos and a reputation for being good. These aspects of trust are far removed from the conceptualization of HMI trust as something almost indistinguishable from simple *reliance*, and it clearly shows why ethos is crucial for understating HMI trust defined in ways similar to HHI trust.

Trust is a relational phenomenon – something existing in each person in a relationship (Sica & Sætra, 2023). Trust is established "by creating a psychological connection with the audience", and if this connection fosters the perception of practical wisdom, character or virtue, and goodwill, "the trustful connection is made" (Jamar, 2001), which links directly with ethos. This means that trust and trustworthiness should not be confused with simple reliance and reliability. However, even if the concept of trust is broadly construed, it captures only *parts* of ethos. According to Weresh (2012), trust "is a persuasive, source-relational attribute of ethos", and ethos can thus be seen as a broader concept *explaining* trust, reliance, *and* other aspects relevant to HMI.

On way to conceptualize trust is to see it as based on evaluations of *ability*, *benevolence*, and *integrity* (Sica & Sætra, 2023). Before the latter two become relevant, the trustor must consider the trustee *able* to perform some action or behaviour – otherwise trust becomes irrelevant. For example, if a 5-year old says that they will help me finish a task at hand, related to reporting on a research project to a funding body, I know that the child is incapable of doing so and me *trusting* the child to deliver on the promise never becomes relevant. Likewise, if an intelligent dishwasher is equipped with a GPT-like conversation module, and states that it will join and help me climb Mount Everest, the obvious lack of *ability* means that we never enter the domain of trust. Such empty promises will, however, clearly influence ethos-relevant aspects such as perceived intelligence, character and benevolence, as the machine either does not understand it's capabilities or quite simply lies. However, when the trustor sees the trustee as competent (to reduce their uncertainty and behave in ways that might allow the trustor to reach their goals), considerations of whether they *will* act on this competence will naturally follow, and this is based on evaluations of integrity and benevolence.

While ethos and trust are clearly not the same, the linkages and relevance of understanding ethos to understand trust is clear. An orator's *expertise* and *intelligence* affect their ability-based trust and the degree to which we can *rely* on the orator. Their *character* affects integrity-based trust, and their goodwill affects benevolence-based trust. However, ethos is broader than trust, and understanding someone's ethos allows us to more effectively understand the degree to which they are trusted. Ethos can be seen as a universal and constant concept, whereas *trust* is usually defined as something that becomes relevant once, for example, some degree of risk and interdependence emerges. Ethos is consequently seen as a key explanatory variable for trust, and exploring machine ethos will enrich the understanding of trust in HMI.

3. Machine ethos and human-machine trust

With the preceding background in place, we are well positioned to

ask: Does a machine have ethos? And, if so, what kind and how can we conceptualize machine ethos? In this section the notion of machine ethos is further developed, following the work of Brahnam (2009), who argues that ethos is an important topic for AI researchers and that it provides a useful theoretical framework for understanding design and ethical issues involved in machine's persuasive power, believability, credibility, and trustworthiness.

Despite this apparent relevance, few researchers in HMI have mobilized the concept of ethos to explore the relationships between humans and machines. Even in books on human-machine *communication*, ethos is almost completely absent (Guzman, 2018; Guzman et al., 2023). One single exception exists in these books – a chapter on the rhetoric of social robots, in which Fritz (2018) relates how ethos is important for social robots, as also recognised by the companies developing robots such as the social robot *Jibo*. Jibo's creators see ethos as instrumental in their design and marketing strategies, and have developed a set of "value statements" to guide developers and engineers, describing how Jibo, for example, is curious, strives to belong, needs his family, etc. (Fritz, 2018).

Much research on HMI refers to and applies the "computers-are-social-actors" (CASA) paradigm (Nass et al., 1994), which implies a recognition and appreciation of the extent to which humans respond to various machines *as if* they were social actors akin to humans (Sætra, 2020a). It's important to emphasize that CASA does not entail a commitment regarding whether machines are *actually* social actors – simply that they are perceived as such. However, the very fact that they are perceived as such has important consequences for how humans behave, and it will consequently have real effects. While it might be true that humans perceive machines as social actors, this need not have any implications for the moral or legal status of machines – it can be seen as a simple descriptive theory of human psychology.

While CASA refers to computers, the paradigm is also used for machines in general and, for example, robots (Dautenhahn, 1999; Edwards et al., 2019). Much CASA research relies on taking theories from HHI and testing the relevance to HMI, and this is also amenable to drawing on research on ethos and testing the relevance to machines.

However, I approach machines not as straight forward social actors, but as *compound* and *perceived* social actors clearly different from human social actors. Their compound nature refers to how they are in part autonomous agents and perceived as such, while they are simultaneously representatives of and perceived as products developed and deployed by organizations and networks of human beings. Ethos provides a useful framework for exploring the relationship between the perceived ethos of machines, which will be based on the perception of the machine itself, *and*, for example, the reputation of the organizations behind them. Machine ethos will consequently also be compound, or *mixed*. The autonomy of machines – and machine ethos – might be little more than a mirage caused by humans' deeply social nature and desire to attribute meaning to our surroundings and interactions, and a "veil of complexity" introduced with the complication of modern technology (Sætra, 2021a). Nevertheless, the mirage *matters*, and is consequently highly relevant both for understanding how to design machines for effective HMI and for understanding and avoiding negative consequences related to how humans perceive and respond to machines.

I will first develop and describe this model of analysing the machine ethos of compound social actors. Secondly, I'll discuss how robots could have an introductory ethos. Thirdly, I move on to some considerations related to how ethos is built – and shaped – during interaction.

3.1. A model of machine ethos

> *But in machine-generated messages, especially those produced by conversational agents,* who *is it that is speaking? Where is ethos when the human is radically removed and a machine is the one producing the speech (Brahnam, 2009, p. 27)?*

Machines, and in particular the highly anthropomorphic autonomous machines such as social robots, are different from human orators in nature. Before going into to the details of the kind of ethos they can have and build I will here present the theoretical model that guides the exploration of machine ethos.

First, a machine can be perceived as having a character of their own, much like a human orator. In the above quote Brahnam (2009) does not mean that machines have in fact become fully autonomous, but that machine action is *seemingly* detached from human intervention, and that users might interact with a machine in situations where the human aspect of the machine is fully obfuscated for the user. In addition, a machine can be perceived as a *medium* – an entity through which human developers, designers, executives, etc. intentionally and accidently communicate (Sætra, 2021b). This means that there is an additional indirect source of ethos to consider, as humans interacting with machines might in varying degrees interpret and evaluate the ethos of the individuals and/or the organization responsible for the machine in addition to any ethos the machine itself might be perceived to have.

The theory of *polyphony* in rhetoric is useful for highlighting how machines can simultaneously have several layers of ethos – or several "voices" – as they are in one sense an agent, and in another sense the conveyor of the messages theoretically attributable to other human beings (Amossy, 2001; Sætra, 2021a). The machine will simultaneously and at various times be perceived as a *thing* and something akin to a *person*, and this complicates the development and analysis of a stable machine ethos and should urge us to also consider their potential category as something else – some new "other" (Gunkel, 2023).

The main model of machine ethos is shown in Fig. 2. This shows how machine ethos is *perceived* by the user as a result of a combination of the machine and its origin and control entities. The *makers* are relevant, as their reputations and activities related to user-communication, support, etc., influence machine ethos. The machine itself is connected to, but also clearly distinct from, maker ethos, and its reputation and appearance will often be the dominant source of machine ethos. However, the influence of the two sources will vary greatly with machine type, corporation, and users. Finally, the *situated* user and their knowledge, attitudes and interactions with the machine is a key determinant of machine ethos. The premise here is that one cannot assume that "the user" is some universal abstract being. Users will differ along a range of dimensions, and identifying and adapting to these is crucial for building machine ethos. Machine ethos, then, is *contextual, compound,* and *relational.*

3.2. Introductory machine ethos

The "pre-communicative opinion" of the entity is here referred to as introductory ethos, but it is a phenomenon with many other names as well, such as initial, extrinsic, and situated ethos (Amossy, 2001; Andersen & Clevenger, 1963; Brahnam, 2009; Giffin, 1967). The main idea is that introductory ethos is formed and shaped based on input preceding direct interaction.

For example, when ChatGPT took the world by storm late 2022, word of mouth and mainstream news outlets quickly generated a *public image* of the service, and very few subsequently encountered the system blindly. The images users had developed varied, of course, as did the public evaluations. The popularity of the service makes it reasonable to assume that people had a preconception about the system as competent, and surprisingly so, which relates to the perceived expertise of ChatGPT. However, many will also have heard stories about how it tends to "hallucinate", which in turn might affect perceived character or reliability, but potentially also perceived expertise. There will also be varying degrees of perceived goodwill, as some see the company behind the system (OpenAI) as a problematic part of the world of "big tech" (Sætra et al., 2021). Users and regulators have questioned how ChatGPT uses personal data and issues related to privacy, leading Italian regulators to ban the service in April 2024 (McCallum, 2023), and this might

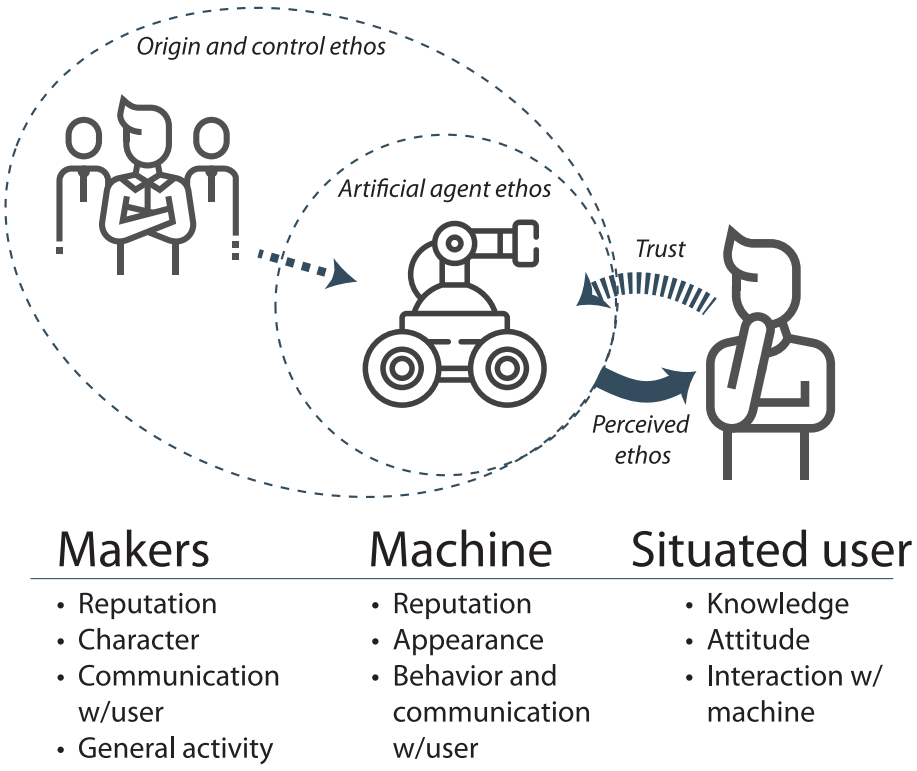

Fig. 2. Main model of machine ethos.

influence the perceived character and benevolence of OpenAI. Simultaneously, stories also showed that the system tended to promote, for example, American values and particular political ideologies (Baum & Villasenor, 2023), which might lead some to perceive the system as poorly aligned with their own interests. All this relates to the system's *introductory* ethos and contributes to shaping users' perception of the system as they start interacting with it, or quite simply refuse or abstain from interacting with it based on these preconceptions.

Regarding the three elements of machine ethos, intelligence or expertise will mainly – but not exclusively – be considered an aspect of the *machine*, while character and goodwill to a larger degree is seen as a product of both the machine and the wider network of responsibility and supporting organizations. Benevolence, in particular, will often be linked to user perceptions related to the motives and "goodness" of the company behind the machine – particularly when this link is strong and known. However, anthropomorphism will also make this a relevant factor for the machine itself, and for machines that are not tightly linked to a producer. For example, Sony will likely be more prominent in user's evaluations of *Aibo* than SoftBank is for users of *Nao*, if we consider a general audience. If we then imagine more maligned companies, such as Musk's Tesla or Meta releasing a robot, it seems likely that company ethos will heavily influence machine ethos and user trust in these robots.

3.2.1. *Machine characteristics*

One source of introductory ethos are various forms of "introductions" – formal or informal and based on text, word-of-mouth, videos, etc. Much quantitative research has used *introductions* to texts, for example, as the intervention in attempts to measure ethos (Andersen & Clevenger, 1963). For a machine, we must here consider a wide range of sources serving as possible introductions, such as official marketing, reviews, news stories, informal posts containing descriptions, images, or videos of the machine in social media, or quite simply how people have heard about a machine in talk with friends, for example. This relates to the *reputation* of a machine, which is shown to influence audience beliefs (Brahnam, 2009).

Research has shown that introductions matter, and the conveyed images of *expertise*, *authoritativeness, position, perceived willingness to see several sides to issues,* etc. have complicated but real effects (Andersen & Clevenger, 1963). For a machine, the level of *expertise* and *reliability* (partly character) can clearly be influenced before interaction. Some of these aspects can be actively managed by the developers of a machine, while others are beyond their control. Nevertheless, trying to understand the broad "introduction" users have had to the machine will be crucial for understanding its introductory ethos, and this will also allow them to adjust and tailor the official introduction in ways to combat negative and promote desired perceptions. However, as we have seen, this will require differentiated communication and efforts aimed at different audiences, as cultures vary significantly regarding both how they perceive machines and what characteristics and virtues are respected. For example, while robots tend to be seen as *things* in Western culture, they are much more heavily embedded in social contexts in Japan (Gunkel, 2018; Robertson, 2018), and such differences will have significant consequences for understanding machine ethos.

While it is not possible to fully demarcate the reputation of a machine from its producer, it is likely that the audience will have different – although linked – perceptions of the two entities. When considering introductory ethos, it will be useful to bring in existing knowledge of the corporation's ethos, such as user surveys related to brand perception, user experience, etc. Different machines also require different types of ethos for effective use. For example, a toy robot will benefit from having a different ethos than a machine aimed at being used in life-support situations in the healthcare sector. How the machine is made will partly influence this ethos, but so will the corporation's ethos. For example, *Hasbro* might not be too successful in marketing a life-support machine, while their ethos *would* support them in efforts to market a toy robot.

I have already mentioned how different machines have varying strength of ties to their parent companies, and whenever there is a mismatch between company ethos and the ethos that would benefit the machine, it will be important to obfuscate or minimize the corporate

branding of the machine, and instead focus on attempts to market the machine as its own brand, or even an autonomous agent. For example, if ByteDance, the Chinese company behind *TikTok* were to launch a social robot in the US or Europe, they might consider it strategically important to launch the robot under a new brand or at least to minimize brand linkage to avoid negative connotations and user fears that, for example, the robot will collect personal data and share it with actors the users don't trust or like.

If we consider the machine a *medium* or *message*, then the corporation becomes the *author* of said machine (Sætra, 2021b). While we have seen from rhetorical studies that the message or text is never fully autonomous, the world of autonomous machines potentially changes this somewhat, as emphasising autonomy and encouraging anthropomorphism might in fact make the machine appear to be the author, medium, and message at once. While humans *appear* to be radically removed from humans, this does not entail an *actual* loss of, for example, human responsibility and legal liability related to the consequences of the machine's actions, however (Sætra, 2021a).

3.2.2. Audience characteristics

In addition to the network of responsibility for the machine – including the machine itself – various aspects of the audience contribute to shaping a machine's *introductory* ethos. One obvious source of insight into the variables affecting a machine's introductory ethos is the vast literature on the factors that influence technology acceptance by use of the technology-acceptance model (TAM) (Davis, 1985). Reviews of the use of the model has shown how, for example, gender, user types, culture, technology anxiety, experience, etc. influence the users perceptions and acceptance of technology (Lee et al., 2003; Marangunić & Granić, 2015). This research is relevant for identifying and dealing with baselines of different audience types regarding the evaluation of technology, as acceptance is arguably linked to machine ethos.

In addition to the TAM literature, there is a growing literature on how various audience variables affect trust, but also some studies related more explicitly to ethos. With regard to gender, for example, research has also show that audience gender can affect a robot's persuasiveness, for example how Ågren and Thunberg (2022) found that men rated a particular robot's "ethos" higher than women. However, Thellman et al. (2018) reports that *women* found the robots used in their experiment more persuasive than men, so the effect of audience gender is clearly complex and something requiring more detailed analysis. Another example shows how a user's experience with a system influences their preferences for system interaction style, and that, for example, experienced users were shown to prefer that a dialogue-based system did not include self-references (Wenger, 1991).

In addition, I have already discussed at some length how *culture* and the social context is crucial for understanding one's audience, and that it will at times be both easier and less legally and ethically challenging to base a machine's actions on the profile of a group – or society – rather than attempt to individually understand and profile each user. This is an issue I return to in the following.

3.3. Machine ethos through interaction

> Manipulating ethos, within ethical limits, is a powerful persuasive tool (Weresh, 2012, p. 272).

Once interaction with a machine begins, we are in the realm of ethos *management,* through which derived or intrinsic (or discursive or invented) ethos emerges (Amossy, 2001; Andersen & Clevenger, 1963; Brahnam, 2009). The machine's introductory ethos is thus the baseline, and interaction moderates of displaces initial perceptions in a wide variety of ways (Amossy, 2001). If the introductory ethos is favourable for achieving the goal the machines is designed for, it can be highlighted and emphasised. However, when the introductory ethos precludes effective interaction, the machine's various means of modifying ethos – communication, appearance, and behaviour – must be aimed at building a new and different image of the machine. This will be done most effectively if a preceding analysis of its introductory ethos has been performed, so that the machine can actively counter or support the audience's preconception.

In addition to analysing the appearance and behaviour of the machine, the people behind the machine will also be able to act during the interaction in ways that affect derived machine ethos.

3.3.1. How is machine ethos built, changed, shaped, and formed?

Ethos can be built through using a combination of the canons of rhetoric, where much focus is on speech and the use of words, but where appearance and behaviour are also strongly represented. The canons include *inventio* (invention), *dispositio* (arrangement), *elocutio* (style), *memoria* (memory), and *pronuntiatio* (delivery) (Frost, 2005). The most relevant ones for my current undertaking are invention, arrangement, style, and delivery.

Invention refers to the various means through which proof can be discovered (Crowley & Hawhee, 2004). This matters for shaping ethos, as this is where communication and interaction are *determined*, and where a machine constructs its behaviour and responses to the user. One example of how invention matters is the determination of how to construct arguments, as research shows that it is, for example, quite often beneficial to cite sources and refer to authorities, when these are perceived as such by the audience (Andersen & Clevenger, 1963). Furthermore, research shows that users at times discount and disregard computer advice – also when good – but that explanations and proven past performance is conducive to increase trust in and use of machine advice (Alvarado-Valencia & Barrero, 2014).

This will naturally be particularly important for machines that are somehow perceived as unreliable and low on expertise by users, and in such cases the machine must also make sure that their references to sources are *also* believable. Otherwise they risk further undermining their ethos. For example, early popular large-language models, including ChatGPT, *invented* academic sources when prompted by users for citations, and while these were seemingly quite plausible, alert users soon discovered that they were pure fiction.

Including explicit remarks about self also influences ethos. While the effects of different statements will vary based on both machine type and, for example, the degree to which the user anthropomorphises the machine, research has shown that, for example, "including conciliatory remarks, statements of self-praise, obvious attempts at ethical appeal enhances the speaker's status" (Andersen & Clevenger, 1963).

Arrangement, or organization, of the material in the speech also matters for ethos (Weresh, 2012). One example of how to effectively build ethos in certain audiences (particularly those with some education), is the use of a "both sides" approach, where not just the arguments supporting the orator's position are presented, but also *opposing* viewpoints and arguments. The opposite strategy (one-sided) has been shown to be more useful with audiences not having completed high school and those already strongly supporting the orator's views (Andersen & Clevenger, 1963). Arrangement also entails the use of, for example *storytelling* and *narrative*, which can be important for promoting ethos through source-relational ethos and making explicit connections between experiences and knowledge shared by orator and audience (Weresh, 2012).

Ethos is also partly built through ***style***, which here refers not just to what is said and done, but *how* it is said and done (Amossy, 2001). Style is occasionally confused with unnecessary ornament and seen as detrimental to persuasion and a mark of excess (Kallendorf & Kallendorf, 1985). But style and the use of the figures from rhetoric – such as the antithesis, hyperbole, and metaphor – is not about artificially dressing up language. It is about making one's point effectively, and often in simple ways (Kallendorf & Kallendorf, 1985). Style also conveys impressions of, for example, eloquence, wit, culture, and personality. Humour in particular is a stylistic element that is often seen as a potentially

effective but also double-edged tool in the style toolbox (Weresh, 2012). One salient example of how developers actively shape the ethos of machines is found in the new large language-model *Grok* released by Elon Musk's company xAI in Novmber 2023. Unlike its competitors, such as ChatGPT, Grok is designed "with a bit of wit and has a rebellious streak" (xAI, 2023), and drawing on the rhetorical canon of style allows for the analysis of the implications of such design choices on ethos.

Regarding metaphors, the effective use of them allows the orator to create a bond with the audience and draw upon their existing knowledge and beliefs, simultaneously fostering persuasion *and* the bond required for ethos (Weresh, 2012), further highlighting the need to understand the context of discourse. Style is, quite simply, about "conveying the proper image", and when business people dress according to the clients their meeting, and choose their words and phrasing in ways that suit their audience, we are in the realm of ethos-building style (Kallendorf & Kallendorf, 1985). In rhetoric, style tends to refer to words, but in the context of HMI, aspects related to appearance and body language, for example, must also be seen as style elements that influence ethos. This means that ethos is not just *verbal* but fully multi-modal, as it is conceived in research on source credibility (Giffin, 1967; Kallendorf & Kallendorf, 1985).

Finally, **delivery**, including the use of voice, gestures, etc, has clear effects in human persuasive efforts, and research has also shown that a robot's voice, for example, influences the audience's attitude to the robot (Edwards et al., 2019; Wenger, 1991). Voice is particularly important for machines that communicate verbally, as research has consistently shown that people automatically infer various personality and physical characteristics based on an orator's voice, and they can to some degree identify *class differences* in orators only by hearing them speak (Andersen & Clevenger, 1963). While a machine could be argued to *have* no class, it's voice and choice of phrasing, etc., will most likely give the users an impression of class, meaning that this will be important for creating – or preventing – identification and the perception of joint interests and beliefs, for example. Simpler computer interfaces could also be explored through the canon of delivery, as aspects such as colours, icons, buttons, and other objects becomes part of the rhetoric of HMI (Wenger, 1991).

In addition to this, a machine's physical appearance will likely have consequences for ethos, as for example *dress* and *manner* are shown to affect the attitude of an audience towards a speaker (Andersen & Clevenger, 1963; Giffin, 1967). This highlights the need to carefully consider machine appearance and design, and also to deal with problematic issues related to, for example, perceived racialization of machines. Giffin (1967), for example, refers to how race and other "irrelevant characteristics" is shown to have "real importance in interpersonal communication situations", and that some studies show that race influences the audience acceptance of a message. This is also explored through *ethos* and the challenges faced by black female authors (Pittman, 2006). Another aspect discussed in the literature on robot persuasiveness is how a machine's gender can influence ethos. Surveying some of the literature indicating that robot gender *could* influence persuasiveness, neither Thellman et al. (2018) nor Ågren and Thunberg (2022) found evidence for the effects of robot gender. Others have explored *avatar* gender – or perceived lack thereof through androgynous features – and the implications of credibility, and have found that it *has* effects on credibility (Nowak & Rauh, 2008).

Furthermore, machines are *staged* in various ways, and research on believability from media studies also show how aspects such as camera position, lighting, scene composition, and other accidental factors matter (Brahnam, 2009). These considerations are particularly relevant to machines who interact with humans through *virtual* agents or interfaces in which such aspects can easily be manipulated. Several studies have shown how, for example, *interactivity* has significant effects on on user perception of machines and their credibility (Brahnam, 2009; Burgoon et al., 2000; Johnson & Kaye, 2016).

A key design choice also relates to anthropomorphising features in the machine's presentation and interface. Entirely removing human qualities is now largely seen as impossible, as user's will anthropomorphise just about anything (Brahnam, 2009; Reeves & Nass, 1996), but some consider it a goal to *reduce* and counteract excessive anthropomorphisation (Shneiderman et al., 2018). Machines are by definition embodied, but they might also contain interfaces with anthropomorphic features, and it is also possible to add various anthropomorphic cues to the machine itself – such as eyes and facial expressions (Brahnam, 2009).

3.3.2. The role of the human-machine relationship and bond

> The discourse will have no effect if the audience is misconstrued and bears no resemblance to the empirical addressees (Amossy, 2001, p. 6).

As stressed above, successfully building ethos requires actively drawing on and utilising knowledge and beliefs that are perceived as *common* to the machine and human (Amossy, 2001). To successfully build and manage ethos, this requires the machine – or surrounding systems – to monitor, analyse, store, and be able to retrieve information about the social context, and potentially also the individual user. The latter can be problematic both for legal and ethical reasons, as the profiling of individuals through storing and analysing personal data is heavily restricted in contexts such as the EU where GDPR applies. Utilising data about groups and cultures could help prevent legal trouble, but it could still be ethically problematic to profile users based on knowledge of their "type" or others like them (Sætra, 2020b).

We might also note that while monitoring and getting to know the user will be crucial for effectively linking beliefs, knowledge claims, and preferences, this process – if conducted in a way that the user perceives as the machine being curious about them – could also generate the impression in the user that the machine is *interested* in them, and this could independently help raise the impression of goodwill and alignment of interests (Braet, 1992). This kind of behaviour could even be decoupled from the actual *storing* and use of the data for profiling, but the user would naturally start to wonder about the strength of the machine's interest in them if it did not remember what it was told or shown.

This raises the question of whether it is sufficient that a machine gathers and relates its interaction to the user's knowledge and beliefs, or if it is also important that the human perceive these to be *shared*. The latter would require that the machine is perceived as having its own knowledge – which is perhaps relatively straightforward – but also *beliefs*, which could suggest something of a human quality. Even more problematic is the use of and reference to common *values*. Does a machine have values? Can it? Is it possible to generate a situation in which a human in HMI feels that *we* – the human and machine – share certain values?

Regardless of objections related to the nature of machines, as I return to in 4.2, I note that machine *learning* in the literal sense will be crucial for developing and shaping machine ethos. Learning from interaction will let the machine adapt appropriately to context and create a bond and relationship with the user. Interestingly, the notion of a machine that behaves and learns can also be related to the etymology of ethos, which is often related to *habit* and *custom* (Corts, 1968). This could be likened to the process in which a learning machine through interaction with others picks up on norms and customs and develops a set of *habits* conducive to the development of the desired ethos (Halloran, 1982). Developing machine ethos, thus, can at least in part be done through a process of *habituation*.

3.3.3. Corporate activity in interaction

Finally, we must briefly consider how corporate activity during the interaction phase can influence derived ethos. For example, if during the use of a machine produced by fictional company *Z,* the company is exposed for having poor cybersecurity, for selling customer data to

others, or some other scandal, machine ethos could change even if the machine remains exactly the same as it was before. The user's evaluation of integrity and benevolence, in particular, might suffer. This also relates to the above comments about challenges related to, for example, the GDPR and ethical issues related to the profiling of individuals.

In addition, the company can actively engage with the customer through a range of activities, such as general public marketing, but also personalized newsletters to owners and users of the machine. More indirect activities, such as efforts aimed at increasing the corporation's image through efforts aimed at corporate social responsibility (CSR) or sustainability, could also help shape company ethos, and thus machine ethos.

The activities of the corporation could be aimed at changing the perception of *corporate* ethos (e.g. CSR activities), but it could also be aimed at communicating and shaping *machine* ethos more directly. For example, a newsletter could describe and explain lesser-known aspects of the machine's workings and functionality, increasing perceived expertise.

Finally, the corporation will be able to directly affect and change the machine through, for example, service, repairs, and updates both to hardware and software. This means that while the machine can be designed to be dynamic, learning, and responsive, it will *also* be liable to change fundamentally during the interaction phase. One example of how a product changing during the interaction phase leads to likely changes in ethos and user perceptions was seen in the radical changes introduced to the *Replika* relationship chatbot in 2023. The social partner app suddenly acquired a new and changed personality, and user reactions led the company developing the app to reverse some changes an re-introduce the capability for "erotic roleplay" (Tong, 2023).

3.3.4. Key considerations for machine design

Considerations of derived ethos allow designers and engineers to make better informed design choices, guided by the kind of impression they want to give and what kind of relationship they want the machine to develop with the user. This includes choices related to appearance and voice, as we have seen that colour and gender, for example, matters, and also that voice and sound is important factors influencing ethos.

More importantly, the need for *learning* and bonding with users based on this learning is central for good ethos. This could be related to existing approaches to *impression management* (Bozeman & Kacmar, 1997; Sætra, 2023c), which requires the same. When informed by the theoretical framework for machine ethos, it provides a rich source of both theory and philosophical and empirical research. Doing so also requires extensive knowledge of – and an adaptive approach to – users, and how the machines might be used in different contexts and by different users over time, which will also change what sort of behaviour is conducive to good ethos.

When managing machine ethos, it is also important to remember that a machine is *not* human, and that balancing the attempts to create *likeness* and identification to promote goodwill, for example, can easily be perceived as creepy and wrong – "contrived camaraderie" (Weresh, 2012, p. 270) – and backfire. Designers are also in a transparency dilemma when managing ethos, as they must consider the consequences of emphasising *machine* nature (transparency and honesty) and downplaying the same to encourage anthropomorphism (to give the impression of likeness and familiarity).

4. Discussion: the implications of machine ethos on trust

Having discussed some of the main aspects related to machine ethos, what remains is consider the relevance of and objections to the concept.

I will first describe why the concept is useful and instrumental in the analysis and design of machines in general, before showing how a focus on machine ethos might be coupled with the postmodernist critique of ethos in an effort to challenge unjust and discriminatory social structures and institutions, and to question existing power relations. Lastly, I present the main objections to the idea of machine ethos, focusing both on philosophical objections and objections based on the potential negative ethical consequences related to various uses of the concept.

4.1. The relevance of machine ethos

What, then, are the conclusions regarding machine ethos and trust arrived upon after the preceding considerations? In this section I describe how researchers, designers, developers, engineers, regulators, and others, could improve the understanding and design of machines by approaching human-machine interaction at least in part through the lens of machine ethos.

4.1.1. Machine ethos as the foundation for trust, reliance, and credibility in HMI

I first ask, like Brahnam (2009), whether it "is possible for [artificial] agents to inspire confidence and trust?" More specifically, whether a consideration of *ethos* helps us understand the nature of HMI trust. Machine ethos can be argued to be useful due to the depth of the theoretical framework it provides for approaching central HMI phenomena such as trust, reliability, and credibility. It also provides a way to approach the balancing act related to optimal levels of anthropomorphisation. Importantly, it helps understand some of the dangers related to over-anthropomorphisation. These could be imaged as "disasters in credibility" once the machine's charade, so to speak, is broken (Brahnam, 2009), as "insincerity, if revealed, has disastrous consequences" (Weresh, 2012). Using ethos manipulatively – and this relates to various approaches used by engineers to make machines relatable – can easily backfire, and learning from *human* failures in manipulative persuasion could be useful for HMI as well. Machine ethos as a lens for HMI might in fact lead to a situation where machines are made *less* human-like to prevent issues like overtrust and to better calibrate user perceptions with the capabilities and goals of machines.

Research on trust in HMI could also often benefit from broader considerations of ethos. Studies such as those examining *trust* in automated leadership – algorithmic management – already explore parts of the antecedents of trust, as for example Höddinghaus et al. (2021) discusses benevolence, integrity, and ability as elements of trustworthiness. Relating this to machine ethos would arguably give the authors a broader and more consistent arsenal of explanatory variables.

4.1.2. Understanding anthropomorphism and humans' perceptions of machines

Machine ethos also opens the door to explorations of the notion of a *rhetorical contract* in HMI. Wenger (1991), for example, explores "the construction of a user's social identity in interaction with a computer", and this line of inquiry follows naturally from analyses of machine ethos and the social contexts and structures that determine the formation of machine ethos in HMI.

One important benefit of analysing machine ethos is that it allows for a consideration of the machine as a kind-of social actor while also being mindful of the *separate* source of machine ethos connected to the company or the people responsible for the robot. Research has, for example, shown how trust and prosocial behaviour between individuals are affected by the introduction of *third-parties* in the interaction (Spadaro et al., 2023), and the company can be construed as such a third-party. The company is necessarily a part of HMI, as they have made the choices about how the machine appears and behaves, but they are also actively involved in providing updates for and service of the machine, and they will often also be known to have access and control over the data gathered and processed. The company is consequently in a control and monitoring position, and while the users will likely at times forget or ignore this, it will at other times, and for some people, be very important for determining machine ethos and trust. Machines, then – just like humans – are never isolated entities.

It is also worth noting that understanding machine ethos could be of

relevance not just for analysing autonomous machines, but also machines used for computer-mediated communication (CMC). The use of telepresence robots and avatars in various computer-mediated settings would here be of relevance.

4.1.3. Machine ethos opening for challenging the doxa of machines and power relations

> … extant social systems inhere normative values that maintain injustice and inequity and that largely determine and legitimate who and what gets valued, who gets silenced, and who gets to speak (Holiday, 2009, p. 392).

An interesting aspect of the postmodern critique is that it alludes to a "democracy of texts", where the orator's "products" are treated as independent social constructs just as much the property of the reader as of the author, and in which "the reader's views are also as good as the opinions proffered by the author herself" (Brahnam, 2009, p. 18). This allows us to challenge established authority and doxa, and it also allows us to see why an orator and their products will be evaluated very different in different time periods – naturally and rightly so. Ethos is consequently quite useful for identifying and challenging doxa, as "[t]o have *ethos* is to manifest the virtues most valued by the culture to and for which on speaks" (Halloran, 1982). A concern related to manipulating ethos is the temptation to perpetuate biases by automatically or cynically resorting to problematic "stock structures or other share knowledge structures" (Weresh, 2012).

Designers, then, must make a choice alluded to above. Do they a) accept this doxa and make their machines to please, or b) do they use their knowledge of doxa to challenge it? This question relates inquiries of ethos to broader questions of design justice (Costanza-Chock, 2020), and general inquiries into representation in the design process, *representation* in and by technology, and potentially the normativity of machines such as social robots (Sætra et al., 2022). Inclusion, Holiday (2009) argues, is a key ethical goal of postmodernism, and this relates both to inclusion in the sense of who gets authority and agency through their social positions, and for our purposes also who is included in the design and development of the machines that come to co-inhabit our societies and take part in the joint invention of reality (Holiday, 2009).

Feminist scholars have also highlighted the complex and reflexive relationship among ethos and *politics* (Holiday, 2009). This perspective underscores the way "rhetoric both *invents* and *is invented* by humans, individually and collectively" (Holiday, 2009, p. 390), which further highlights the need for rhetors and scholars of rhetoric to take *responsibility* for their ethos (Reynolds, 1993). This involves both taking responsibility for how their rhetorical activity influences politics, but also that it is necessary for orators to actively engage with and disclose their own situated nature and various sources of authority. This will, for example, entail providing the audience with an account of *where* one is from – one's background and locatedness (Reynolds, 1993).

Doxa, then, is not just something to take into account when shaping machine ethos, but something we all create through our activities. It is also arguably *arbitrary* (Bourdieu, 2000), and something we can and should challenge and reshape if existing doxa reinforce and reproduce structures of injustice. Machine ethos as a concept can help us uncover and identify problematic rhetorical activities, and thus promote reflection amongst designers and developers about how the machines they make influence the world in which they are deployed.

4.2. Objections to the notion of machine ethos

Despite the purported relevance of machine ethos, there are two key objections to the concept that must be considered. One is based on the idea that machines are entities of a kind that makes talk of ethos – which is to some a *human* quality – inapplicable to them. The other is based on consideration related to how the use of the concept of machine ethos could have a range of negative ethical implications, and that, in short, machine ethos might be *unethical*.

4.2.1. Objection 1: machines cannot have ethos

> The 'available means' – which, let it be added, may prove incommensurate with the intention to persuade – cannot be specified or programmed in advance (Sellars, 2006, p. 59).

One major objection to any talk of "machine ethos" could be based on the premise that machines themselves are *never* more than the medium through which human being acts (Sætra, 2020a; 2021a). Some refer to the lack of credibility "as the human body is removed from discourse", leading to a situation in which humans in HMI effectively interacts with *no-body* – interactions without human partners directly present (Brahnam, 2009). One might also refer to the lack of *face* and that there is no *other* present in the interaction (Sætra, 2020a), and this could certainly make it difficult to speak of machine ethos. In addition, Aristotle stressed how ethos is always the result of the orator's voluntary or deliberate choice (Pittman, 2006; Sattler, 1947), which raises the question of whether machines are capable of such, or if they simply manifest the behaviour programmed in them by other humans. If so, their behaviour might not reveal their nature or character, so to speak. Finally, some researchers' emphasis on *situated* humans and *lived* experience (Reynolds, 1993), will potentially make it seem absurd to speak of the ethos of a *machine*.

There are several counter-objections to this objection. First, we might argue that what matters is not whether or not humans are *actually* present in the interaction, and that it suffices that users *perceive* machines as human-like enough (Turkle, 2017). In addition, I have shown that humans are also actually present in the interaction through their part in the broader network of responsibility where the machine is only one part, and designer, developers, and executives, etc. are also present (Sætra, 2021a).

Furthermore, many now argue that it is possible to construe machines as social entities with sufficient status to become *partners* in the interaction, often based on the *relationships* formed between humans and machines (Coeckelbergh, 2010). Others speak of the potential for robots to be morally relevant *others* and that they can have "face" in the Levinasian sense (Gunkel, 2018, 2023). Such approaches would clearly also make it highly relevant to explore the ethos of machines, and this second approach would also potentially decrease the importance of examining the compound ethos proposed in this article, and more directly severing the tie between producers and machines.

4.2.2. Objection 2: machines should not have ethos

Regardless of whether or not machines can be made and programmed in ways that makes the concept theoretically useful, others might argue that promoting machine ethos is ethically problematic for a number of reasons.

First, machine ethos could promote the idea that machines need to build detailed profile of individuals in order to maximize the potential for aligning their interactions with the preferences and attitudes of the user. This could be seen as problematic in itself, as privacy could be seen as an intrinsic good that is valuable *even if* the information is not abused. Observation and the monitoring of individuals could be seen as a violation of a right to privacy, but it could also be seen as a form of *interference* that necessarily changes the individuals' behaviour and is inimical to individual liberty (Sætra, 2019a).

Second, using personality profiles built either on the collection of personal user data or assumed or known likeness with other people entails risks of user manipulation or other forms of influence which is also detrimental to the interests and liberty of the user (Sætra, 2019b, 2020b; Sætra & Mills, 2021; Véliz, 2020). When the machine has a lot of information about the proclivities and interests of the user, and this is coupled with knowledge of how to most effectively persuade and make

use of rhetoric or for example nudging, the machine and corporation behind it could be argued to hold too much power over individuals.

Third, some might argue that talk of machine ethos entails running errands for big tech companies that would be interested in having their machines appear to be human-like, and that this might even lead to a situation where company responsibility and liability for machines are challenged and obfuscated (Birhane & van Dijk, 2020; Sætra, 2021a). In addition to the potential for a shift in legal liability, some might fear that increased attention to the idea of robot moral status is detrimental to *humans'* moral status (Sætra & Fosch-Villaronga, 2021). One line of argument could entail that moral consideration is a zero-sum game, and that any increased moral consideration of machines would come at the expense of humans. For example, if machines are likened to the subjugated minorities of the past that have gone from morally inconsiderate to morally considerate – for example slaves, women, animals – we could find ourselves in a situation where considerations about machines leads us to remake "values, virtues, and standards" and create an "alternative model of ethos" (Pittman, 2006). While some will balk at the notion that *machines* could be seen as another "minority", this is exactly the kinds of questions being asked by some in the robot ethics community (Gunkel, 2023).

A fourth and more general objection, which is linked to the preceding ones, relates to the potential that designers discover the value of ethos and use this to make increasingly anthropomorphic machines that seek to maximize positive ethos to improve user trust. This relates to ethical challenges related to, for example, robot deception and robot betrayal (Danaher, 2020; Sætra, 2021c), as emphasised in the robot ethics literature. While this could surely promote effective HMI, Brahnam (2009, p. 36) provides the following suggestion:

> Rather than foster the suspension of disbelief in an attempt to create a separate imaginary being, developers should open the channels to reality testing and build character from that exchange. They should acknowledge the fact that agents are not human and strive to make the human agencies standing behind the agents transparent.

All the preceding objections represent vital considerations for robot ethics and ethics in HMI, and any use of machine ethos must be coupled with an eye to how the concept could potentially be abused and also have negative consequences even when used with good intentions.

However, Brahnam's warning and suggestion is fully in line with a *descriptive* use of machine ethos established in this paper. While the concept could be used to create machines that encourage anthropomorphisation, it could also be used to better understand the negative consequences of poorly aligned user perceptions and machine capabilities.

5. Conclusion

> Ethos is character. Character implicates trust. Trust is based on relationships. Relationship persuades (Weresh, 2012, p. 229).

In this article, I have shown that the concept of *machine ethos* is relevant for understanding some of the key concepts discussed with different terminologies in HMI. 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.

However, machine ethos is more than an alternative to these concepts, as it allows researchers to draw on a broader theoretical concept that arguably precedes and *determines*, for example, trust. Even if the more specific phenomena are studied, then, machine ethos will be important to consider as a background phenomenon.

While machine ethos is tightly linked to ethos as it is used in HHI, it is also clearly different. The compound character of machine ethos and the dual nature of machines as both autonomous and *mediums* for others requires careful considerations of both users and machines *in context* to understand machine ethos. Part of the analyses required entails asking questions related to the degree to which users will perceive various machines as having intelligence, character, and goodwill. For machines that are heavily anthropomorphised – as a consequence both of design and various inclinations for different users – all three aspects remain relevant for the *core* machine ethos directly stemming from the machine. For other machines, all three aspects remain relevant, but the role of the humans behind the machine – it's *authors* of sorts – will also be quite important in many cases. This helps link the analysis of machine, user, *and* corporations in the complete ecosystem of machine ethos.

I have also discussed, but partly intentionally sidestepped, the deeper philosophical question related to the "real" nature of machines, and questions related to whether a machine can *actually* have character and benevolence, for example. The questions related to the social, moral, and legal status of machines are indeed relevant to explorations of machine ethos, but settling these questions are not necessary here, as anthropomorphisation and user *perception* of such features in machine suffices to make machine ethos a meaningful concept – even if some might argue that it *should not* be. However, seeing machines as entities capable of having an ethos will have consequences, and I have discussed both positive consequences related to improved understanding of the relationship between user perceptions and machine capabilities and various ethically problematic consequences.

Either way, the social nature of HMI is emphasised in machine ethos, as ethos is formed in the "intersection between speaker or writer and listener or reader" (LeFevre, 1987). As shown – it makes sense to see ethos as arising in interaction, but it is not a *shared* phenomenon per se. Rather, it is a series of mirror-interactions in which the audience has a perception of the orator, the orator has a perception of the audience, and through interaction these perceptions that need not be connected to anything "real" shape the relationship to the benefit or detriment of achieving the goals of HMI.

One major benefit of using the broad concept of ethos for understanding machine behaviour and HMI is that it forces the analyses of social structures and the position of different groups and individuals. While this *could* be used to manipulate machine ethos and users, we might also hope that such analyses leads to a recognition of the existence of injustice and problems related to existing doxa, and that such a realization leads to the emerge of an experienced ethical responsibility to challenge such problems (Holiday, 2009).

Credit author statement

Henrik Skaug Sætra: All aspects and processes involved in conceptualizing and writing the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Ågren, I., & Thunberg, S. (2022). Robot persuasiveness depending on user gender. *Proceedings of the 10th International Conference on Human-Agent Interaction*.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems, 21* (4), 12–17.

Alvarado-Valencia, J. A., & Barrero, L. H. (2014). Reliance, trust and heuristics in judgmental forecasting. *Computers in Human Behavior, 36*, 102–113. https://doi.org/10.1016/j.chb.2014.03.047

Amossy, R. (2001). Ethos at the crossroads of disciplines: Rhetoric, pragmatics, sociology. *Poetics Today, 22*(1), 1–23. https://doi.org/10.1215/03335372-22-1-1

Andersen, K., & Clevenger, T., Jr. (1963). A summary of experimental research in ethos. *Communication Monographs, 30*(2), 59–78. https://doi.org/10.1080/03637756309375361

Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4). https://doi.org/10.1609/aimag.v28i4.2065, 15-15.

Anderson, M., & Anderson, S. L. (2011). *Machine ethics*. Cambridge University Press.

Aristotle. (2018). *Rhetoric*. Hackett Publishing Company, Inc.

Aroyo, A. M., de Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., Jones, S., Lutz, C., Sætra, H. S., Solberg, M., & Tamò-Larrieux, A. (2021). Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn. Journal of Behavioral Robotics, 12*, 423–436. https://doi.org/10.1515/pjbr-2021-0029

Baum, J., & Villasenor, J. (2023). *The politics of AI: ChatGPT and political bias*. https://www.brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/.

Beldad, A., De Jong, M., & Steehouder, M. (2010). How shall I trust the faceless and the intangible? A literature review on the antecedents of online trust. *Computers in Human Behavior, 26*(5), 857–869. https://doi.org/10.1016/j.chb.2010.03.013

Birhane, A., & van Dijk, J. (2020). Robot rights? Let's talk about human welfare instead. Proceedings of the AAAI/ACM conference on AI, ethics, and society. In *Language and symbolic power*. Harvard University Press. Bourdieu, P. (1991).

Bourdieu, P. (1991). *Language and symbolic power*. Harvard University Press.

Bourdieu, P. (2000). The historicity of reason. In P. Bourdieu (Ed.), *Pascalian meditations*. Standford University Press.

Bozeman, D. P., & Kacmar, K. M. (1997). A cybernetic model of impression management processes in organizations. *Organizational Behavior and Human Decision Processes, 69*(1), 9–30.

Braet, A. C. (1992). Ethos, pathos and logos in aristotle's rhetoric: A re-examination. *Argumentation, 6*, 307–320. https://doi.org/10.1007/BF00154696

Brahnam, S. (2009). Building character for artificial conversational agents: Ethos, ethics, believability, and credibility. *PsychNology Journal, 7*(1).

Brudner, A. (2007). *Constitutional goods*. Oxford University Press.

Burgoon, J. K., Bonito, J. A., Bengtsson, B., Cederberg, C., Lundeberg, M., & Allspach, L. (2000). Interactivity in human–computer interaction: A study of credibility, understanding, and influence. *Computers in Human Behavior, 16*(6), 553–574. https://doi.org/10.1016/S0747-5632(00)00029-7

Campbell, C. P. (1998). Rhetorical ethos. In N. Susanne, C. P. Campbell, & R. Driven (Eds.), *The cultural context in business communication* (pp. 31–47). John Benjamins Publishing Company.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology, 12*(3), 209–221.

Corts, T. E. (1968). The derivation of ethos. *Speech Monographs, 35*(2). https://doi.org/10.1080/03637756809375583

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Crowley, S., & Hawhee, D. (2004). *Ancient rhetorics for contemporary students*. Pearson Longman.

Damasio, A. (2006). *Descartes' error: Emotion, reason, and the human brain*. Quill.

Danaher, J. (2020). Robot betrayal: A guide to the ethics of robotic deception. *Ethics and Information Technology*, 1–12. https://doi.org/10.1007/s10676-019-09520-3

Dautenhahn, K. (1999). Robots as social actors: Aurora and the case of autism. *Proc. CT99*. The Third International Cognitive Technology Conference, August, San Francisco, Davis, F. D.. *A technology acceptance model for empirically testing new end-user information systems: Theory and results* Massachusetts Institute of Technology].

Davis, F. D. (1985). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. Doctoral dissertation: Massachusetts Institute of Technology.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*(3–4), 177–190.

Edwards, C., Edwards, A., Stoll, B., Lin, X., & Massey, N. (2019). Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions. *Computers in Human Behavior, 90*, 357–362. https://doi.org/10.1016/j.chb.2018.08.027

Fritz, L. (2018). Child or product? The rhetoric of social robots. In A. L. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 67–82). Peter Lang.

Frost, M. H. (2005). *Introduction to classical legal rhetoric: A lost heritage*. Routledge.

Giffin, K. (1967). The contribution of studies of source credibility to a theory of interpersonal trust in the communication process. *Psychological Bulletin, 68*(2), 104. https://doi.org/10.1037/h0024833

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *The Academy of Management Annals, 14*(2), 627–660.

Gunkel, D. J. (2018). *Robot rights*. MIT Press.

Gunkel, D. J. (2023). *Person, thing, robot: A moral and legal ontology for the 21st century and beyond*. MIT Press.

Guzman, A. L. (Ed.). (2018). *Human-machine communication: Rethinking communication, technology, and ourselves*. Peter Lang).

Guzman, A. L., McEwen, R., & Jones, S. (Eds.). (2023). *The SAGE handbook of human-machine communication*. Sage.

Halloran, S. M. (1982). Aristotle's concept of ethos, or if not his somebody else's. *Rhetoric Review, 1*(1), 58–63. https://doi.org/10.1080/07350198209359037

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Höddinghaus, M., Sondern, D., & Hertel, G. (2021). The automation of leadership functions: Would people trust decision algorithms? *Computers in Human Behavior, 116*, Article 106635. https://doi.org/10.1016/j.chb.2020.106635

Holiday, J. (2009). [ter]vention: Locating rhetoric's ethos. *Rhetoric Review, 28*(4), 388–405. https://doi.org/10.1080/07350190903185049

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion*. Yale University Press.

Isaksson, M., & Flyvholm Jørgensen, P. E. (2010). Communicating corporate ethos on the web: The self-presentation of PR agencies. *Journal of Business Communication, 47*(2), 119–140. https://doi.org/10.1177/0021943610364516, 1973.

Jamar, S. D. (2001). Aristotle teaches persuasion: The psychic connection. *The Scribes Journal of Legal Writing, 8*, 61.

Johnson, T. J., & Kaye, B. K. (2016). Some like it lots: The influence of interactivity and reliance on credibility. *Computers in Human Behavior, 61*, 136–145. https://doi.org/10.1016/j.chb.2016.03.012

Kallendorf, C., & Kallendorf, C. (1985). The figures of speech, ethos, and Aristotle: Notes toward a rhetoric of business communication. *Journal of Business Communication, 22*(1), 35–50. https://doi.org/10.1177/002194368502200102, 1973.

Lee, Y., Kozar, K. A., & Larsen, K. R. (2003). The technology acceptance model: Past, present, and future. *Communications of the Association for Information Systems, 12*(1), 50. https://doi.org/10.17705/1CAIS.01250

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

LeFevre, K. B. (1987). *Invention as a social act*. Southern Illinois UP.

Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes, 126*, 88–106.

Marangunić, N., & Granić, A. (2015). Technology acceptance model: A literature review from 1986 to 2013. *Universal Access in the Information Society, 14*, 81–95. https://doi.org/10.1007/s10209-014-0348-1

McCallum, S. (2023). *ChatGPT banned in Italy over privacy concerns*. BBC. https://www.bbc.com/news/technology-65139406.

McCormack, K. C. (2014). Ethos, pathos, and logos: The benefits of Aristotelian rhetoric in the courtroom. *Wash. U. Jurisprudence Rev., 7*, 131.

McCroskey, J. C. (1966). *Scales for the measurement of ethos*. https://doi.org/10.1080/03637756609375482

Merton, R. K. (1972). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*(11), 1905–1922. https://doi.org/10.1080/00140139408964957

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI conference on Human factors in computing systems*.

Nowak, K. L., & Rauh, C. (2008). *Choose your "buddy icon" carefully: The influence of avatar androgyny, anthropomorphism and credibility in online interactions. Computers in Human Behavior, 24*(4), 1473–1493. https://doi.org/10.1016/j.chb.2007.05.005

Offerdal, T. S., Just, S. N., & Ihlen, O. (2021). Public ethos in the pandemic rhetorical situation: Strategies for building trust in authorities' risk communication. *Journal of International Crisis and Risk Communication Research, 4*(2), 247–270. https://doi.org/10.30658/jicrcr.4.2.3

Perelman, C. (1989). *Rhétoriques. Editions de l'Université de Bruxelles*.

Pilsch, A. (2018). *The Ethos of Mr. Robot. Present Tense, 7*(1).

Pinkard, T. (1986). Freedom and social categories in hegel's ethics. *Philosophy and Phenomenological Research, 47*(2). https://doi.org/10.2307/2107437

Pittman, C. (2006). Black women writers and the trouble with ethos: Harriet jacobs, billie Holiday, and sister souljah. *Rhetoric Society Quarterly, 37*(1), 43–70. https://doi.org/10.1080/02773940600860074

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press.

Reynolds, N. (1993). Ethos as location: New sites for understanding discursive authority. *Rhetoric Review, 11*(2), 325–338. https://doi.org/10.1080/07350199309389009

Robertson, J. (2018). *Robo sapiens japanicus: Robots, gender, family, and the Japanese nation*. Univ of California Press.

Sætra, H. S. (2019a). Freedom under the gaze of Big Brother: Preparing the grounds for a liberal defence of privacy in the era of Big Data. *Technology in Society, 58*, Article 101160.

Sætra, H. S. (2019b). When nudge comes to shove: Liberty and nudging in the era of big data. *Technology in Society, 59*, Article 101130. https://doi.org/10.1016/j.techsoc.2019.04.006

Sætra, H. S. (2020a). The parasitic nature of social AI: Sharing minds with the mindless. *Integrative Psychological and Behavioral Science, 54*, 308–326. https://doi.org/10.1007/s12124-020-09523-6

Sætra, H. S. (2020b). Privacy as an aggregate public good. *Technology in Society, 63*, Article 101422. https://doi.org/10.1016/j.techsoc.2020.101422

Sætra, H. S. (2021a). Confounding complexity of machine action: A hobbesian account of machine responsibility. *International Journal of Technoethics, 12*(1). https://doi.org/10.4018/IJT.20210101.oa1

Sætra, H. S. (2021b). Robotomorphy: Becoming our creations. *AI and Ethics, 2*, 5–13. https://doi.org/10.1007/s43681-021-00092-x

Sætra, H. S. (2021c). Social robot deception and the culture of trust. *Paladyn. Journal of Behavioral Robotics, 12*(1). https://doi.org/10.1515/pjbr-2021-0021

Sætra, H. S. (2023a). Challenges for the inclusion of robots in social institutions. In *Social robots in social institutions* (pp. 589–594). IOS Press.

Sætra, H. S. (2023b). The hidden assumptions of variable-based social science. In J. Valsiner (Ed.), *Farewell to variables*. Information Age Publishing.

Sætra, H. S. (2023c). *Machiavelli for robots: Strategic robot failure, deception, and trust roman 2023*. Busan, Korea.

Sætra, H. S., Coeckelbergh, M., & Danaher, J. (2021). The AI ethicist's dilemma: Fighting big tech by supporting big tech. *AI and Ethics, 2*, 15–27. https://doi.org/10.1007/s43681-021-00123-7

Sætra, H. S., & Danaher, J. (2022). *To each technology its own ethics: The problem of ethical proliferation*. Philosophy & Technology.

Sætra, H. S., & Ese, J. (2023). Shinigami eyes and social media labeling as a technology for self-care. In H. S. Sætra (Ed.), *Technology and sustainble development: The promise and pittfals of techno-solutionism* (pp. 53–69). Routledge. https://doi.org/10.1201/9781003325086-2.

Sætra, H. S., & Fosch-Villaronga, E. (2021). Research in AI has implications for society: How do we respond? *Morals & Machines, 1*(1), 60–73.

Sætra, H. S., & Mills, S. (2021). Psychological force, liberty and technology. *Technology in Society*. , Article 101973. https://doi.org/10.1016/j.techsoc.2022.101973, 69.

Sætra, H. S., Nordahl-Hansen, A., Fosch-Villaronga, E., & Dahl, C. (2022). *Normativity assumptions in the design and application of social robots for autistic children*. TBA. *submitted to SHI conference*.

Sattler, W. M. (1947). Conceptions of ethos in ancient rhetoric. *Communication Monographs, 14*(1–2), 55–65. https://doi.org/10.1080/03637754709374925

Segal, J., & Richardson, A. W. (2003). Introduction. Scientific ethos: Authority, authorship, and trust in the sciences. *Configurations, 11*(2), 137–144.

Sellars, R. (2006). *Rhetoric. Theory, Culture & Society, 23*(2–3), 59–60.

Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., & Diakopoulos, N. (2018). *Designing the user interface: Strategies for effective human-computer interaction* (6 ed.). Pearson Education Limited.

Sica, A., & Sætra, H. S. (2023). In *Technology we trust! But should we? Human-computer interaction*. HCII 2023, Denmark.

Spadaro, G., Molho, C., Van Prooijen, J.-W., Romano, A., Mosso, C. O., & Van Lange, P. A. (2023). Corrupt third parties undermine trust and prosocial behaviour between people. *Nature Human Behaviour, 7*(1), 46–54. https://doi.org/10.1038/s41562-022-01457-w

Tavani, H. T. (2015). Levels of trust in the context of machine ethics. *Philosophy & Technology, 28*, 75–90. https://doi.org/10.1007/s13347-014-0165-8

Thellman, S., Hagman, W., Jonsson, E., Nilsson, L., Samuelsson, E., Simonsson, C., Skönvall, J., Westin, A., & Silvervarg, A. (2018). He is not more persuasive than her: No gender biases toward robots giving speeches. In *Proceedings of the 18th international conference on intelligent virtual agents*.

Tong, A. (2023). AI chatbot company Replika restores erotic roleplay for some users. *Yahoo! Finance*. https://finance.yahoo.com/news/ai-chatbot-company-replika-restores-184630336.html.

Turkle, S. (2017). *Alone together: Why we expect more from technology and less from each other*. Hachette UK.

Véliz, C. (2020). *Privacy is power*. Bantam Press.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. W. H. Freeman and Company.

Wenger, M. J. (1991). On the rhetorical contract in human—computer interaction. *Computers in Human Behavior, 7*(4), 245–262. https://doi.org/10.1016/0747-5632(91)90013-Q

Weresh, M. H. (2012). Morality, trust, and illusion: Ethos as relationship. *Legal Comm. & Rhetoric: JAWLD, 9*, 229. xAI. (2023). *Announcing Grok*. Retrieved November 7 from https://x.ai.

xAI. (2023). *Announcing Grok*. Retrieved November 7 from https://x.ai.