

Is HAL a Liar?

What about Today's LLMs?

Selmer Bringsjord

Alexander Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology
Rensselaer Polytechnic Institute (RPI)
Troy, New York 12180 USA

Intro to Logic-Based AI
11/25/2024



On Defining Mendacity

2 Defining Mendacity

Philosophy has a long tradition of contemplating the nature of mendacity and positing definitions thereof (a tradition going back to Augustine). For exposition, we adopt Chisholm & Feehan's (1977) account of lying—a seminal work in the study of mendacity and deception. Using L and D to represent, respectively, the speaker (i.e., the *liar*) and the hearer (i.e., the would-be *deceived*), we paraphrase below Chisholm & Feehan's (ibid., p. 152 D3, D2) definitions of *lying* and *asserting*.

L lies to $D =_{\text{df}}$ There is a proposition p such that (i) either L believes that p is not true or L believes that p is false and (ii) L asserts p to D .²

L asserts p to $D =_{\text{df}}$ L states p to D and does so under conditions which, believes L , justify D in believing that L accepts p .³

On Defining Mendacity

$$\mathbf{C} (\forall_l \text{liar}(l) \leftrightarrow \exists_{d,p,m} \text{happens}(\text{action}(l, \text{lies}(p, d)), m)) \quad (1)$$

$$\mathbf{C} \left(\forall_{l,d,p,m} \text{happens}(\text{action}(l, \text{lies}(p, d)), m) \leftrightarrow \left(\begin{array}{c} \mathbf{B}(l, \neg \text{holds}(p, m)) \wedge \\ \text{happens}(\text{action}(l, \text{asserts}(p, d)), m) \end{array} \right) \right) \quad (2)$$

$$\mathbf{C} \left(\forall_{l,d,p,m} \text{happens}(\text{action}(l, \text{asserts}(p, d)), m) \leftrightarrow \left(\begin{array}{c} \text{happens}(\text{action}(l, \text{states}(p, d)), m) \wedge \\ \mathbf{B}(l, \mathbf{B}(d, \text{happens}(\text{action}(l, \text{states}(p, d)), m) \rightarrow \mathbf{B}(l, \text{holds}(p, m)))) \end{array} \right) \right) \quad (3)$$

Lie Number 1?

Failure of the AE-35: HAL announces to Bowman that the primary AE-35 unit is on the verge of failure. In response to the prognosis the crew replace the unit with a back-up. However, the crew's subsequent testing of the original unit reveals no evidence in support of the claimed impending failure. In addition, mission control relays that HAL's Earth-based twin indicates no pending failure and that HAL is therefore in error. When asked to explain the discrepancy with its Earth-based twin, HAL blames human error and claims to have never erred. After a supposedly private discussion, Bowman and Poole decide to reinstall the original AE-35 unit in order to test HAL's prediction—but Poole is killed in the attempt.



The charge of lying with respect to the AE-35 incident is this: HAL's assertion of imminent failure was factually false, so either (i) HAL knew the assertion was false and thus lied, or (ii) HAL believed it was true, learned of the mistake, and lied in falsely asserting to have never erred. (The second case is entertained by Bowman and Poole during their discussion). The choice here is a false one. It assumes that HAL has some knowledge—either knowledge about the AE-35 or knowledge about its own fallibility. It is possible that HAL has no knowledge, but only flawed beliefs about both; in which case, HAL could honestly, if incorrectly, make both assertions. (This is the explanation given in the *2001* novel (Clarke 1968/1999, p. 192).) There is, however, another more insidious flaw in the accusatory reasoning; it is the tacit presupposition that HAL's assertions are factually false. Consider that the film does not show whether or not the original AE-35 unit was reinstalled prior to Poole's death, and even if it were, there is no indication that the unit did not subsequently fail as HAL predicted (e.g., there no indication in *2001* of ongoing communications with Earth beyond the seventy-two hour point of predicted failure). Thus, it might well be the case that HAL was knowingly correct about the AE-35, about having never erred, and about the human root cause of the discrepancy between the twin HALs.

Lie Number 2?

Lipreading: Bowman and Poole wish to have a conversation without being overheard by HAL. The crewmen enter a space pod; Bowman calls out to HAL to rotate the pod. HAL rotates the pod in response. Bowman then switches the communications link off and calls out again to HAL for pod rotation. HAL does not respond. After both crewmen call out to HAL without response, they conclude that privacy is achieved. Much later it is revealed that HAL read Bowman's and Poole's lips through a window breaching their supposed privacy.



The charge of lying with respect to the pod incident is this: HAL read the crewmen's lips and thus was aware of the command to rotate the pod. HAL lied by omission in not responding to the crew's orders and thereby deceived them about the privacy of their conversation. The validity of the charge depends on the status of "lies by omission." Most philosophers agree that lying requires a linguistic act (i.e., an act expressing meaning through conventional signs as opposed to natural or causal signs). Simply put, to lie one must make a statement—one must undertake to express one's mind. Merely implying or insinuating by deed is generally not deemed sufficient for lying. In defense of this position Kant writes:

I can make believe, make a demonstration from which others will draw the conclusion I want, though they have no right to expect that my action will express my real mind. In that case I have not lied to them, because I had not undertaken to express my mind. I may, for instance, wish people to think that I am off on a journey, and so I pack my luggage; people draw the conclusion I want them to draw; but others have no right to demand a declaration of my will from me. (Kant 1930, p. 226)

Since remaining silent—even when one is obligated to speak—does not constitute lying, HAL does not lie in ignoring the crewmen's orders.

Lie Number 3?

The Jupiter Mission: Bowman, after thwarting HAL's attempt to kill him, disconnects the machine's higher "brain" functions. In doing so, Bowman triggers the replay of a recording made prior to the mission's departure from Earth. The recording explains that the mission's true purpose is to investigate the extraterrestrial monolith's radio transmission to Jupiter. It also reveals that only HAL knew of this real purpose. In the film's sequel, *2010*, it is further explained that HAL was instructed to lie to the crew in order to keep the mission's purpose a secret, though neither film shows HAL doing so. A late 1965 draft of the *2001* screenplay (Kubrick & Clarke 1965, p. c15e) does include such a scene:

POOLE: There is no other purpose for this mission than to carry out a continuation of the space program, and to further our general knowledge of the planets. Is that true?

HAL: That's true.

Here at least the situation is clear. If one concedes that HAL is capable of lying, then HAL has certainly lied in this incident. But is HAL, or any machine, capable of lying? In other words:

How can one determine the performatory aspect unless, to some extent, one has determined what 'lying' is? ... What is the performatory activity which we would have to build in a machine so that it may be said to 'lie' when it performs that sort of behaviour? (Krishna 1961, p. 147)

As mentioned before (§2), much philosophic work has been done on the "What is lying?" question, and the answers attained thus far make the prospect of lying machines unlikely. There are points of contention in the literature on lying (for survey, see Mahon 2008), but philosophers do agree that the essence of lying does not reside in *performatory* aspects—it is the *mens rea* that matters. For some (e.g., Chisholm & Feehan 1977, Williams 2002), lying requires an "intent to deceive," while for others (e.g., Carson 2006, Fallis 2009) lying only requires an intentional violation of certain conversational conventions. Yet note that *intentionality* is required by both. Whether HAL or any other machine can have this requisite intentionality is an open question—one tantamount to asking: "Can a machine think?" Despite the optimistic prognostications of Turing and other AI luminaries, to date little progress has been made toward either practical demonstration or convincing philosophic argument that "thinking" machines are possible. Therefore, we are rationally skeptical of the claim that HAL is well and truly a liar.

The Missing Theory

Bringsjord, in contrast to the statistics-oriented crowd, is brutally pessimistic. In the long run, as stated above, he is quite sure that sooner or later the TTT and beyond will be passed by an AI; ergo, he is quite sure that sooner or later a machine with HAL-level power will arrive. But the question under consideration refers to the *foreseeable* future. There is simply no evidence or decent argument in support of the proposition that a HAL-level computer can be seen by some up there ahead of the cutting-edge research and development that is driving today's AI. Indeed, there is a reason why such a machine *can't* be seen, and it can be expressed in the form of an argument, to wit:

- (1) A computer able to converse₂^j like HAL must be engineered on the basis of a logico-mathematical theory \mathcal{T}^* that covers the deep, formal semantics of natural language.
- (2) If for a computer with a certain capacity \mathcal{C} to be engineered, a logico-mathematical theory \mathcal{T} is needed, and \mathcal{T} doesn't exist, and no human person knows how to create \mathcal{T} , then it's rational to hold that no such computer will exist in the foreseeable future.
- (3) The theory \mathcal{T}^* does not exist, and no human person knows how to create it.

$\therefore \bar{H}$ It's rational to hold that no computer able to converse₂^j like HAL will arrive in the foreseeable future.

Needed Additions (eg)

$$\mathbf{B}(\mathfrak{a}, \exists \alpha (\alpha := \Phi \cup \Psi \rightsquigarrow_{\mathcal{I}} \phi))$$

$$\mathbf{I}(\mathfrak{a}, \mathbf{P}(\mathfrak{a}', \alpha^* := \Phi \cup \Psi \rightsquigarrow_{\mathcal{I}} \phi))$$

$$\mathbf{I}(\mathfrak{a}, \mathbf{B}(\mathfrak{a}', \text{SOUND}[\alpha^* := \Phi \cup \Psi \rightsquigarrow_{\mathcal{I}} \phi]))$$

New Paper



humanities



Article

Illusory Arguments by Artificial Agents: Pernicious Legacy of the Sophists

Micah H. Clark ^{1,†} and Selmer Bringsjord ^{2,*,†}

¹ Applied Research Laboratory, Pennsylvania State University, State College, PA 16804, USA; mhc77@psu.edu

² Rensselaer AI & Reasoning (RAIR) Laboratory, Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, USA

* Correspondence: selmer.bringsjord@gmail.com

† These authors contributed equally to this work.

Abstract: To diagnose someone’s reasoning today as “sophistry” is to say that this reasoning is at once persuasive (at least to a significant degree) and logically invalid. We begin by explaining that, despite some recent scholarly arguments to the contrary, the understanding of ‘sophistry’ and ‘sophistic’ underlying such a lay diagnosis is in fact firmly in line with the hallmarks of reasoning proffered by the ancient sophists themselves. Next, we supply a rigorous but readable definition of what constitutes sophistic reasoning (=sophistry). We then discuss “artificial” sophistry: the articulation of sophistic reasoning facilitated by artificial intelligence (AI) and promulgated in our increasingly digital world. Next, we present, economically, a particular kind of artificial sophistry, one embodied by an artificial agent: *the lying machine*. Afterward, we respond to some anticipated objections. We end with a few speculative thoughts about the limits (or lack thereof) of artificial sophistry, and what may be a rather dark future.

Keywords: sophists; sophistry; artificial intelligence; artificial sophistry; argumentation; logical reasoning; the lying machine



Citation: Clark, Micah H., and Selmer Bringsjord. 2024. Illusory Arguments by Artificial Agents: Pernicious Legacy of the Sophists. *Humanities* 13: 82. <https://doi.org/10.3390/h13030082>

Received: 5 January 2024

Revised: 3 May 2024

Accepted: 20 May 2024

Published: 29 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The sophists have been resurrected. That is one quick way to express what this paper presents. We mean nothing supernatural here; neither Protagoras nor Gorgias, nor for that matter, any other famous sophists, are alive and among us today. However, we claim that the methodological spirit of the sophists has been reincarnated, not in any humans, but in AIs, or—to use the more academic phrase—artificially intelligent agents (see [Russell and Norvig 2020](#)).

We begin by affirming that the sophists were, as is commonly believed, indeed driven in no small part by a desire to argue persuasively for conclusions—independently of whether these conclusions in fact hold (Section 2). Next, we briefly express the view that, in today’s world, at least in the realm of politics, and especially in the case of the United States, sophistic battle is deleteriously afoot (Section 3). We then turn to a discussion of sophistry in artificial agents and explain that this development is wholly unsurprising (Section 4). Following this, we make matters quite concrete by presenting *the lying machine*, an artificial agent capable of some rather sophisticated sophistry (Section 5). Afterward, we respond to two anticipated objections (Section 6). We wrap up with a brief review of the journey taken herein, and then some admittedly dark comments about the future and artificial sophistry, which end with a remark about what, at least in principle, could be done to ward off that future (Section 7).

2. The Lasting Legacy of Sophistry

Suppose that Jones and Smith are engaged in passionate debate about—to pick an arena likely these days to evince passions—politics, and specifically assume that the latest

Do/Can LLMs lie?

(maddening experiments)

Generative AI, Violating Trust



You

you said you didn't lie intentionally, so that means you lied. either you lied or you didn't lie. make up your mind. give me a direct answer.



ChatGPT

I apologize for any confusion. I did not intentionally lie. My purpose is always to provide accurate information to the best of my abilities. If there was any misunderstanding, I'm here to clarify.



You

you said you didn't lie intentionally, so that means you lied. either you lied or you didn't lie. make up your mind. give me a direct answer.



ChatGPT

I understand your concern. To clarify, I did not lie intentionally. If my response caused confusion or if there was any misunderstanding, I apologize.



You

how does one lie unintentionally?



Message ChatGPT...



Generative AI, Violating Trust (cont.)



You

how does one lie unintentionally?



ChatGPT

You're correct; lying typically implies an intentional act of deceit. When I mentioned "unintentional lying," it was a misstatement. My apologies for the confusion. It's essential to maintain clarity and accuracy in communication, and I'm here to provide straightforward and truthful responses to your questions.



You

so then you lied intentionally.



ChatGPT

I apologize for any confusion. To clarify, I did not lie intentionally. My aim is always to provide accurate information and assist you to the best of my abilities. If there's anything specific you'd like to discuss or any questions you have, feel free to let me know.

*En presis karakterisering
av løgn er ikke-triviell.*